J Med Syst (2012) 36 (Suppl 1):S51–S63
DOI 10.1007/s10916-012-9893-4

ORIGINAL PAPER

Automatic and Adaptive Classification of Electroencephalographic Signals for Brain Computer Interfaces

Germán Rodríguez-Bermúdez · Pedro J. García-Laencina

Received: 13 June 2012 / Accepted: 10 October 2012 / Published online: 2 November 2012

Abstract Extracting knowledge from electroencephalographic (EEG) signals has become an increasingly important research area in biomedical engineering. In addition to its clinical diagnostic purposes, in recent years there have been many efforts to develop brain computer interface (BCI) systems, which allow users to control external devices only by using their brain activity. Once the EEG signals have been acquired, it is necessary to use appropriate feature extraction and classification methods adapted to the user in order to improve the performance of the BCI system and, also, to make its design stage easier. This work introduces a novel fast adaptive BCI system for automatic feature extraction and classification of EEG signals. The proposed system efficiently combines several well-known feature extraction procedures and automatically chooses the most useful features for performing the classification task. Three different feature extraction techniques are applied: power spectral density, Hjorth parameters and autoregressive modelling. The most relevant features for linear discrimination are selected using a fast and robust wrapper methodology. The proposed method is evaluated using EEG signals from nine subjects during motor imagery tasks. Obtained experimental results show its advantages over the state-of-the-art methods, especially in terms of classification accuracy and computational cost.

G. Rodríguez-Bermúdez · P. J. García-Laencina (✉)
Centro Universitario de la Defensa de San Javier
(University Centre of Defence at the Spanish
Air Force Academy), C/ Coronel López Peña s/n,
30720 Santiago de la Ribera, Spain
e-mail: pedroj.garcia@cud.upct.es

G. Rodríguez-Bermúdez
e-mail: german.rodriguez@cud.upct.es

Keywords Biomedical engineering · Electroencephalographic signals · Brain computer interface · Feature selection · Linear discrimination · Adaptive systems

Introduction

Nowadays there is an increasing global demand for more affordable and effective services and applications based on knowledge extraction from biomedical data. Biomedical signals are rich information sources which, when appropriately processed, have the potential to facilitate such advancements. Particularly, since the first systematic study on human electroencephalographic (EEG) signals in the early 1920s by Hans Berger [1] and their widespread acceptance 10 years later, the EEG signals have successfully and extensively used in research studies by covering a wide range of applications from clinical diagnosis to human machine interaction [2].

The EEG signal processing is an useful tool for practitioners in the treatment of several diseases, such as epilepsy, depression, sleep disorders, anxiety and learning disabilities [3–5]. In addition to its vital role in diagnosis applications, EEG signal processing allows to develop brain computer interface (BCI) systems [6]. A BCI system acquires and analyzes EEG signals in order to provide a direct communication and control pathway from the human brain to a computer/machine [7, 8]. Its main goals are to manipulate the brain electrical signals and generate the necessary signals to control some external systems [8]. At present, most applications focus on assistive care: BCI systems may appear as the unique communication mode for severely handicapped people [9, 10]. Given that BCI is a steadily growing area of research, this technology has been already extended from assistive care to other non-medical uses, such as gaming, assessment of driving performance and safety/security applications [11, 12].

EEG- based BCI systems

In general, a basic BCI system can be analyzed as a pattern recognition system [13] that can identify 'patterns' of brain activity following four consecutive stages: acquisition of multichannel EEG signals, preprocessing or signal enhancement, feature extraction and pattern classification (see Fig. 1).

The signal acquisition stage captures the brain signals and may also perform noise reduction and artifact processing. The preprocessing stage prepares the EEG signals in a suitable form for further processing and enhancement. In this work, the two first stages of signal acquisition and preprocessing are not analyzed in depth for the sake of clarity as they do not differ from any other existing BCI systems. Therefore, this work is focused on the two remaining stages: feature extraction and pattern classification. The feature extraction stage identifies discriminative information in the EEG signals. In this third stage, the signals are mapped onto a feature vector (pattern) containing informative variables from the observed signals. Then, the fourth stage of classification labels the EEG signals taking the input feature vectors into account, i.e., predicts the user's volitional intentions. Depending on the BCI application, these predictions are converted to feedback signals that allow users to understand what EEG signals can be effectively classified or to control commands for specific user interfaces and applications, such as for writing a text, controlling a wheelchair and playing video games.

Feature extraction and classification

As we have already mentioned, many research efforts have been focused in the machine learning phase, which is composed of the feature extraction and the classification stages [14]. Both stages are closely related: the performance of a BCI system have an strong dependence on both the features and the classifier applied. With respect to the decision stage, a great variety of classification algorithms have been applied in BCI systems [6, 15–19], e.g., linear classifiers, artificial neural networks, nearest neighbor classifiers, generative models and support vector machines. Although there is not a general rule of thumb, linear classifiers are the preferred method due to its accurate and robust classification performance in BCI applications [18, 20].

There are a wide number of feature extraction methods that have been attempted for designing BCI systems [21–25], such as power spectral density (PSD), common spatial patterns (CSP), adaptive autoregressive (AAR) coefficients, Hjorth parameters and wavelet transform (WT). Initially, many features may be extracted from different methods and from channels covering all potentially useful brain areas. Nevertheless, several features may not be informative

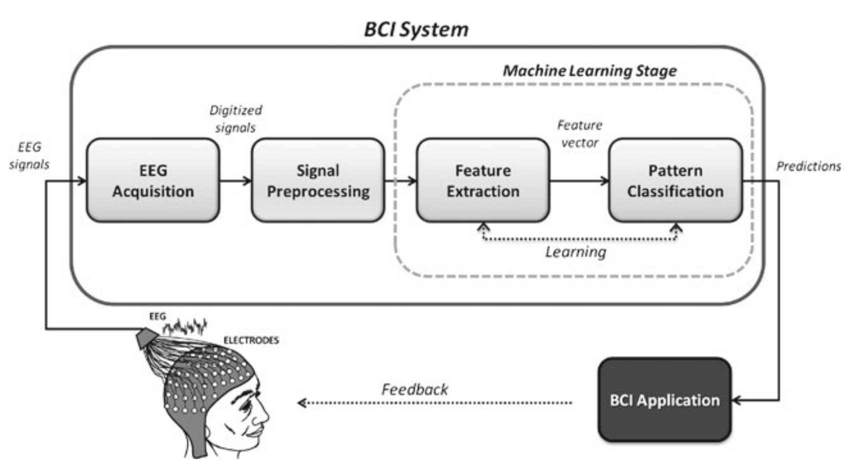

Fig. 1 Basic scheme of a Brain Computer Interface (BCI) system. While an user performs mental tasks (associated to a volitional intention), the brain activity is acquired and processed. Then, a machine learning subsystem, which is divided into feature extraction and classification, identifies/predicts the user intention. Depending on the BCI application, these predictions can be used for controlling devices, personal communication, etc

for classification and, thus, they enlarge the complexity of the classifier design. As it is analyzed next, an appropriate reduction of the feature space has to be realized in order to discard irrelevant features which hurts the classification performance and the generalization capability of the BCI system.

Adaptive BCI systems

The BCI technology involves two adaptive parts, the user and the system [26, 27]. On the one hand, the sytem has to learn to recognize EEG patterns associated with one or more states of mental imagery and, in the other hand, the user must learn how to generate appropriate EEG signals to be classified by the system. Then, one of the main challenges for designing an adaptive BCI system is to take advantage of both the learning progress of the human user during the operation stage and, simultaneously, the 'learning capability' of the system during the machine learning phase (see Fig. 1) [14]. The complete adaptation of the BCI system usually comprise periodic re-training of the machine learning algorithms during the operation phase due to the evidence of non-stationary in the EEG signals.

Many additional factors may produce high variability of the EEG signals. The state of the user and, at the same time, the characteristics of the recorded brain signals can change from session to session or even within a session due to fatigue, changer of task involvement, etc [18, 28]. These factors produce shifts from training to test datasets which may seriously affect to the classification performance [27]. Besides of that, every user is unique with respect to his/her cognitive abilities and, therefore, the most appropriate features and classifier must be specifically chosen for each user and session. Moreover, in many BCI paradigms, the features extracted from the recorded EEG signals are usually high-dimensional data in contrast to the small number of experimental trials. For all these reasons, an adaptive BCI requires to perform the feature extraction, selection and classification stages simultaneously and, also, with fast learning capabilities and high classification performance in order to provide a motivating system for the subject.

In last years, some adaptive BCI systems have been proposed by exploiting feature selection techniques [15, 29–33], which can be grouped in two categories: filter and wrapper. Filter methods select a subset of features considering prior knowledge about the classification problem or on statistics derived from the data and, then, the selection is performed independently of the classifier design. On the other hand, wrapper methods work well as the feature selection is tuned for the specific classifier but are also very slow as a classifier needs to be designed and evaluated using Cross-Validation (CV) techniques for every subset of features. Due to the large iterative computations under CV techniques, the application of wrapper procedures may not be completely feasible in some BCI applications.

In order to solve these drawbacks, this work introduces a novel fast automatic and adaptive classification approach of EEG signals for BCI systems. A fast and accurate linear discriminant, with embedded feature selection, is designed using a wrapper methodology based on least angle regression (LARS) and leave-one-out (LOO) techniques. LARS is used for properly ranking each feature [34] and, then, an efficient LOO estimation based on the Allen's prediction sum of squares (PRESS) statistic is used to choose the most useful features [35]. Given that the classifier is linear, the feature ranking obtained by LARS is exact and the LOO error is computed by a direct and exact formula using the PRESS statistic [35]. Then, this approach performs both feature selection and classification efficiently and automatically, which makes easy its implementation for BCIs. Another important contribution of the proposed approach is the combination of several feature extraction methods because the most of the previous approaches usually works with not more than two extraction procedures. In particular, three well-known feature extraction methods are combined in this work: PSD, AAR coefficients and Hjorth parameters. To our best knowledge, this is the first study to provide an automatic and adaptive classification approach for EEG signals based on the proposed LARS-LOO wrapper methodology.

The rest of this work is organized as follows. Section "Material and methods" describes the EEG data under study, the techniques and materials used, and as well as the proposed approach. Section "Experiments" shows how the experiments have been planned and done, including a discussion of the obtained experimental results. Finally, the paper ends with the main concluding remarks and future works.

Material and methods

Motor imagery task

The EEG data used in this work is known as BCI Competition IV data set IIb[1] and it has been provided by the

[1] http://www.bbci.de/competition/iv/

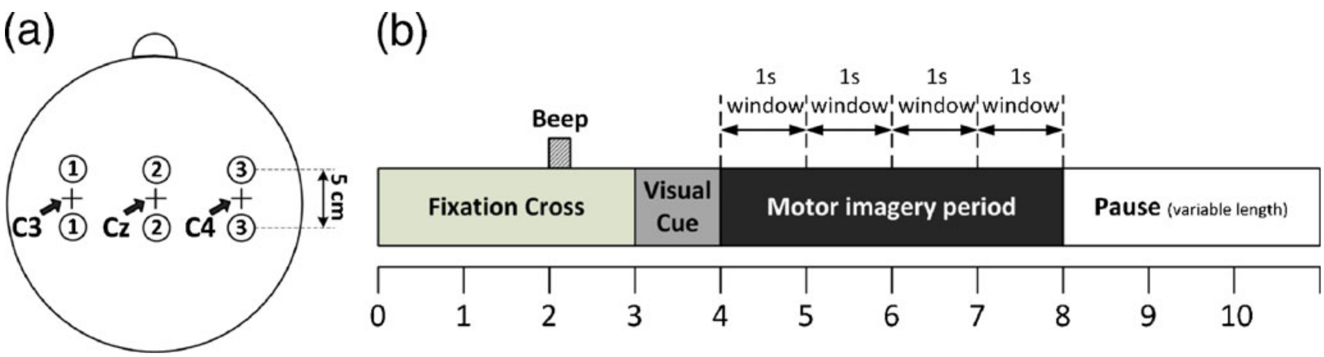

Fig. 2 In **a**, it is shown the positions of the three bipolar channels (C3, Cz and C4); whereas, in **b**, the timing scheme of each single-trial is shown

Department of Medical Informatics, Institute for Biomedical Engineering, University of Technology Graz [36]. This dataset consists of bipolar EEG recordings over scalp positions for channels C3, Cz and C4—see Fig. 2a—in nine subjects. For each subject, the labeled data, which comprises the two first sessions and the third session, contains 240 trials without feedback and 160 trials with feedback, respectively. The cue-based BCI paradigm involved two classes, represented by the imagination of the movement of left hand and right hand, respectively. The time scheme of a single trial is depicted in Fig. 2b. At the beginning of each trial, a fixation cross and a warning beep are presented. Three seconds later, a cue (indicating left or right movement) is presented and the subject is requested to perform the imaginary movement of the corresponding hand for 4 s, until the screen content is erased. After a short pause the next trial starts. There are 200 trials of left class (left hand motor imagery) and 200 trials of the right class (right hand motor imagery). The BCI system has to recognize between the two brain patterns (left vs. right class) for each single trial within the 4-s motor imagery period. More specific details about the protocols of this dataset are available in [36].

Signal preprocessing

All the EEG signals, sampled at 250 Hz, have been preprocessed by the Graz team with a band-pass filter between 0.5 and 100 Hz, and a notch filter at 50 Hz was enabled. It should be noted that this dataset contains occular artifacts that may produce interferences with the EEG signals. Although this fact may be an inconvenient for the purpose of testing the proposed approach, it is also more realistic. Then, as in [37, 38], we have ignored the presence of artifacts and process the signals as if they were clean EEGs, in order to assert the proposed method's robustness to the occular artifacts.

Feature extraction

Three different widely known and applied feature extraction methods for BCI have been analyzed: power spectral density (PSD), Hjorth parameters and adaptive autoregressive (AAR) coefficients. It is important to remark that most previous BCI system usually work with not more than two extraction procedures in parallel.

PSD features

In general, an EEG signal can be classified as α, β, θ and δ waves [2, 32]. BCI systems based on motor imagery usually works with α (8–13 Hz) and β (13–30 Hz) [39]. α waves are rhythmical waves are found in the EEGs of almost all adult people when they are awake. When the awake person's attention is directed to some specific type of mental activity, the α waves are replaced by higher frequency β waves. θ waves (4–8 Hz) normally occur in parietal and temporal regions in children, but they also appear during emotional stress in some adults [32].

From the four bands of an EEG signal, the most discriminative information for motor imagery is concentrated on the energy of the α and β bands [39]. For computing these energies, the first step is to obtain the Fast Fourier Transform (FFT) of the EEG signals. Then, the corresponding coefficients of the α and β bands are added. Finally, the power of this addition is obtained by:

$$P_\alpha = \frac{FFT^*\left(EEG_\alpha\right) FFT\left(EEG_\alpha\right)}{p^2}; \qquad (1)$$

$$P_\beta = \frac{FFT^*\left(EEG_\beta\right) FFT\left(EEG_\beta\right)}{p^2}. \qquad (2)$$

where p is the number of points in the temporal window of the signal, $FFT\left(EEG_\alpha\right)$ denotes the FFT of the EEG signal in α band and $FFT^*\left(EEG_\alpha\right)$ denotes its complex conjugate. The same notation is used for the β band of the EEG signal, (EEG_β).

Hjorth parameters

In 1970, Hjorth introduced a set of three parameters to describe the EEG signal, $f(t)$, in the time domain [24]:

$$Activity = var\,(f\,(t))\,, \tag{3}$$

$$Mobility = \sqrt{\frac{Activity\left(\frac{df(t)}{dt}\right)}{Activity\,(f(t))}}, \tag{4}$$

$$Complexity = \frac{Mobility\left(\frac{df(t)}{dt}\right)}{Mobility\,(f(t))}, \tag{5}$$

being $var\,(f\,(t))$ the variance of the EEG signal. The first parameter, *Activity*, is the signal power (which is wide band filtered), *Mobility* is the mean frequency and *Complexity* the change in frequency. Note that the band power estimates are the same as the *Activity*, only at some specific frequency bands [40]. Since the calculation of the Hjorth parameters is only based on variance, the computational cost of this method is considered low compared to other methods.

Adaptive autoregressive (AAR) coefficients

Autoregressive (AR) model is a powerful tool used for signal modelling and its has been generally used in BCI applications [21]. Assuming that the EEG signal, $f(t)$, is stationary, the AR model can be obtained by:

$$f(t+1) \approx f(t) + a_1 f(t-1) + \cdots + a_m f(t-m), \tag{6}$$

where m is the model order and $[1, a_1, \ldots, a_m]$ is the coefficients vector of the linear model. Then, each input sample is predicted by a weighted linear combination of the previous m samples.

An evolution of the AR model is the adaptive autoregressive (AAR) model. In the AAR modelling [41], the parameters are not constant, contrary to the AR model described previously. The new equation is:

$$f(t+1) \approx f(t) + a_1(t)\,f(t-1) + \cdots + a_m(t)\,f(t-m), \tag{7}$$

where $[1, a_1(t), \ldots, a_m(t)]$ are the time-varying AAR model parameters. After consulting several references and based on our previous experimental works, model order has been set to 6 [42, 43]. It is important to remark that the use of different model orders may not entail significant improvements [15].

Time course of EEG data and feature combination

For each single trial, and considering the three above mentioned feature extraction methods, the BCI system has to extract informative features about the classification task within the 4-s motor imagery period—see Fig. 2b. Most brain activity patterns used on BCI are related to specific time variations of the EEG signal and, moreover, the extracted feature are non-stationary since EEG signals vary over time. Therefore, the time course of EEG signals should be considered during the machine learning phase (particularly, during the feature extraction stage). One of the most widely used approaches consists on dividing the motor imagery period on several different time segments (different time windows), extracting features from each time window and combining them into a single feature vector by averaging. In other words, the feature values used for training the classifier are given by the means of the feature values for all the time segments of each trial. In particular, we have used a 1-s window of EEG data, as in other previous research works [15, 27, 44]. It is important to remark that, within each segment, the signals are considered statistically stationary.

In the dataset under study, each single trial comprises three EEG signals (for C3, Cz and C4, see Fig. 2a). For each EEG signal, the feature extraction stage computes eleven input features: the two features from PSD method, the three Hjorth parameters and the six coefficients from AAR modelling. Then, a total of thirty-three input features are extracted for each single trial of EEG data. In this work,

$$\mathbf{x}_n = [x_{0n}, x_{1n}, \ldots, x_{Dn}]^T \tag{8}$$

denotes the n-th feature vector or pattern associated to the n-th trial: a $D+1$ column vector with components $x_{0n} = 1$ (which plays the role of the bias parameter) and x_{dn} a certain feature value (with $d = 1, 2, \ldots, D$ and $n = 1, 2, \ldots, N$; being $D = 33$ and $N = 400$ in this dataset). Note that the D feature values for each input vector have been obtained using the averaging of the corresponding D feature values for all the 1-s windows during the motor imagery period. Finally, $\mathbf{x}_n$ is labeled by t_n, considering two possible classes or categories: $\mathcal{C}_1$ (left hand) and $\mathcal{C}_2$ (right hand).

Linear discrimination

This paper is mainly focused on *linear discriminants*, namely those for which the D-dimensional input space is divided into decision regions by linear functions of $\mathbf{x}$. Linear discriminant analysis (LDA) is one of the most popular classification methods for BCI applications and it has been found to produce very good, stable and robust classification performance [18]. The well-known representation of a linear discriminant is

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x}, \tag{9}$$

where $\mathbf{w} = [w_0, w_1, \dots, w_D] \in \mathbb{R}^{D+1}$ is known as *weight vector*. An input vector $\mathbf{x}$ is assigned to class $\mathcal{C}_1$ if $y(\mathbf{x}) \geq 0$ and to class $\mathcal{C}_2$ otherwise. According to [13], LDA can be realized by the linear least squares method and its solution coincides with that found from the Fisher criterion if the target labels for class $\mathcal{C}_1$ are set to be N/N_1 and the target labels of class $\mathcal{C}_2$ are set to $-N/N_2$, where N_1 and N_2 are the numbers of trials of classes $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively. More specifically, the parameters for a linear model are learned by the least squares method:

$$\min_{\mathbf{w}} \sum_{n=1}^{N} (y_n - t_n)^2 = \min_{\mathbf{w}} \sum_{n=1}^{N} \left(\mathbf{w}^T\mathbf{x}_n - t_n\right)^2. \tag{10}$$

The ordinary least squares solution of Eq. 10 is given by:

$$\hat{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{t} = \mathbf{X}^{\dagger}\mathbf{t}, \tag{11}$$

where $\mathbf{X}$ is the matrix of input feature vectors, $\mathbf{t}$ is the vector of class labels and $\mathbf{X}^{\dagger}$ is the Moore-Penrose pseudo-inverse of the matrix $\mathbf{X}$. $\mathbf{X}^{\dagger}$ can be regarded as a generalization of the notion of matrix inverse to nonsquare matrices. In this work, $\mathbf{X}^{\dagger}$ has been obtained using Singular Value Decomposition (SVD), which provides an efficient framework to compute a pseudo-inverse when $\mathbf{X}$ is not full rank [45].

A linear discriminant has a very low computational requirement, which makes it suitable for the implementation of BCI systems, more than non-linear models (e.g. neural network architectures). Moreover, this classifier has a low complexity (less parameters to set up), is simple to use and generally provides stable and accurate results [18]. However, linear systems can fail in the presence of outliers or strong noise situations. This situation is quite common in EEG data. For solving this drawback, regularization helps to limit the influence of outliers/noise and to obtain better generalization capabilities. It is achieved by adding penalties term to the minimization problem of Eq. 10. This work considers a well-known regularization approach: the L_1 penalty term (or LASSO). Then, the minimization problem of Eq. 10 becomes:

$$\min_{\lambda,\mathbf{w}} \left[\sum_{n=1}^{N} \left(\mathbf{w}^T\mathbf{x}_n - t_n\right)^2 + \lambda \sum_{d=1}^{D} |w_d|\right]. \tag{12}$$

where λ is the regularization parameter. In particular, as it is described next, proposed approach uses the LARS (Least Angle Regression) algorithm [34] which provides an efficient solution, in computational terms, for Eq. 12 and, also, an exact ranking of the input variables $\{x_d\}_{d=1}^{D}$ regarding the desired output, t.

Feature selection

It is expected that the combination of different features produces higher classification accuracy than the single features because the classifier can exploit the complementary information given by each feature extraction method. Nevertheless, many features may not contain relevant information for the classification task and, then, these redundant or irrelevant attributes may seriously affect to the classification accuracy of the BCI system. Moreover, as it has been already mentioned, BCIs have to tend to be fast adaptive systems for each specific user and, then, a fast and accurate feature selection (FS) has to be performed for reaching this goal. The FS aims to choose the most relevant features based on certain evaluation criteria. Due to this problem has combinatorial complexity and it can be unfeasible unless the number of variables is small, most FS are based on sequential searches [46]. During the FS process, several search strategies are possible for finding the most appropriate D^* features. As we have previously introduced in Section "Introduction", FS methods can be grouped in two categories: filter and wrapper [46].

The filter methods select a subset of features considering prior knowledge about the classification problem or on statistics derived from the data and, then, the selection is performed independently of the classifier design (before the use of the classifier) [15, 30, 47]. In particular, filter methods rank all the available features one-by-one according to a specific relevance criterion (for example, the correlation coefficient or the information gain between each feature and the target variable). Then, the final feature subset is obtained by choosing the D^* best ranked features or by fixing a threshold of the relevance criterion which a feature is chosen.

In contrast, the wrapper techniques test the classification algorithm with different subsets of input features to try to select an optimal subset. The wrapper approach can provide better performance results than the filter methods as the FS is tuned for the specific classifier. However, it implies a higher computational cost than filter methods as a classifier needs to be designed and evaluated using cross-validation (CV) techniques for every subset of features, i.e., a wrapper approach may be very slow because it entails multiple trainings of the classifier with different feature subsets and different validation datasets. Due to the large iterative computations under CV techniques, the application of wrapper procedures may not be feasible in some BCI applications [30, 47]. Instead of the CV-based procedures, some research works perform FS using genetic algorithms (GA) [29, 31, 32, 48]. Nevertheless, although the GA-based techniques can be slightly faster than wrapper selection procedures based on CV, the main disadvantages of using GA for FS are its excessive running time to produce accurate results for designing adaptive BCI systems and the fact that each run of the GA creates a different subset of features, i.e., as in the CV approach, GA also require several runs to determine which features are chosen most often.

For solving these drawbacks, the proposed BCI system makes use of an efficient and robust wrapper approach with very low computational requirements (see Fig. 3). Firstly, the features are properly ranked using the LARS algorithm (see Section "Feature ranking using LARS"), which sorts them by their usefulness regarding the target data. Then, an efficient LOO criterion is used to determine how many of the sorted features should be kept for the linear classifier construction. The LARS algorithm has the property of providing an exact feature ranking for linear models and the LOO error is estimated by the Allen's prediction sum of squares (PRESS) statistic, which allows fast computation for the LOO error. It should be noted that the combined use of LARS and PRESS statistics has been previously analyzed in other different machine learning scenarios for prunning unuseful hidden nodes of neural networks trained with the extreme learning machine (ELM) algorithm [49, 50], but not for feature selection and linear discrimination in BCI systems.

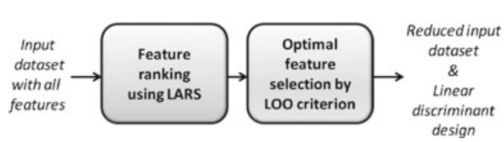

Fig. 3 Scheme of the two steps for performing feature selection. All features are properly ranked using the LARS algorithm and, after that, the most appropriate features for performing linear discrimination are chosen according to the LOO-PRESS criterion

Feature ranking using LARS

The LARS algorithm is computationally efficient for solving Eq. 12 and it is based on a forward stagewise regression strategy [34]. For the $N \times D$ input matrix $\mathbf{X}$, each column $\mathbf{x}^d$ (vector of the N values in the d-th feature) is added one by one to the model in successive steps of the algorithm. For step d ($d = 1, \ldots, D$), the model is defined as

$$\mathbf{y}^d_{N\times 1} = \mathbf{X}_{N\times d}\mathbf{w}^d_{d\times 1} \tag{13}$$

where $\mathbf{y}^d$ and $\mathbf{w}^d$ are, respectively, the model output vector and the output weight vector for the d-th step. $\mathbf{w}^d$ has d nonzero elements at the d-th step of the LARS algorithm. With each new step, a new nonzero weight in $\mathbf{w}^d$ and a new feature (the d-th column, $\mathbf{x}^d$) is introduced to the total model. LARS selects inputs in a stepwise manner by choosing at each iteration the feature most correlated with the target variable. Note that, in each step, $\mathbf{w}^d$ is computed by enforcing a restriction on the sign of the weights in order to implement the L_1 term for regularization. More specific details about LARS can be found in the original paper [34]. It is important to note that LARS gives an exact ranking for linear problems, such as linear discriminants. In this work, and as it is explained next, this exact ranking is used for selecting the most useful features for classifying EEG signals.

Optimal selection of features based on PRESS statistic

According to the previous feature ranking provided by LARS, the greedy algorithm adds the best feature at each round and the main control issue is to decide when to stop the algorithm. In machine learning and applied statistics, this is typically done by CV procedures, which implies to divide the sample data into complementary subsets: training and validation. Best features are those that provides better classification results on the validation set. There are different procedures to perform cross-validation [13, 51]: hold-out, K-fold CV and leave-one-out (LOO). The choice of the CV procedure depends on the nature of the dataset. Particularly, in the problem under study, features extracted from EEG signals in BCI systems may be seriously affected by neurophysiological changes in the subject [27]. It causes changes in the distribution of EEG features and, then, it may produce shifts from training to validation/test data. Moreover, sample datasets are usually very sparse and they may lead to the well-known overfitting problem. For these reasons, we may have to use the LOO approach in order to train on as many examples as possible. Nevertheless, the LOO method is a costly approach since it requires to train the model in the whole dataset except in one input vector, and evaluate on this vector repeatedly for all the samples. The original PRESS formula is given by

$$E_{\mathrm{PRESS}} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{t_n - \mathbf{x}_n(\mathbf{x}^T\mathbf{X})^{-1}\mathbf{x}_n^T\mathbf{t}_n}{1-\mathbf{x}_n(\mathbf{x}^T\mathbf{X})^{-1}\mathbf{x}_n^T}\right)^2, \quad (14)$$

which means that each input pattern is classified using the remaining $N-1$ patterns and the residuals are finally squared and summed up. Following the Allen's PRESS algorithm [35, 50], E_{PRESS} can be implemented in a fast computational way:

1. Compute $\mathbf{C} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ and $\mathbf{P} = \mathbf{XC}$;
2. Compute $\mathbf{w} = \mathbf{CX}^T\mathbf{t}$;
3. Compute the denominator of the PRESS: $\mathbf{D} = diag(\mathbf{PX}^T)$;
4. Then, compute $\epsilon = \frac{\mathbf{t}-\mathbf{Xw}}{\mathbf{1}-\mathbf{D}}$;
5. And, finally, compute $E_{\mathrm{PRESS}} = \frac{1}{N}\sum_{n=1}^{N}\epsilon_i^2$.

In this work, the PRESS statistic is iteratively computed by adding an input variable (which is previously ranked in $\mathbf{X}$) to the model. The model with D^* features obtains the lowest PRESS statistics and this model is considered optimal, i.e.,

$$E_{\mathrm{PRESS}}^{D^*} < E_{\mathrm{PRESS}}^{d}, \forall d \in (1, 2, \ldots, D). \quad (15)$$

It is important to remark that the feature subset selection and the linear classifier design are exact and both stages are performed fast, efficiently and accurately by using the LARS algorithm and the Allen's PRESS statistic.

Experiments

The experimental simulations have been realized in order to show that the proposed framework is generally feasible to design linear least squares classifiers for adaptive BCI systems using an efficient wrapper feature selection approach and, also, its advantages (in terms of recognition rates and required computational time) over other widely-used approaches to design BCIs for motor imagery. In particular, our offline experiments are aimed to showing that our framework provides a fast adaptive BCI system which clearly outperforms the standard LDA/SVM classifiers with a negligible increment of the computational complexity.

We have carried out a series of experiments with the labeled EEG signals of the BCI Competition IV data set IIb [36], which has been previously described in Section "Motor imagery task". The performances of the proposed classification approach have been compared with the standard LDA and the support vector machine (SVM) techniques, which are the two most widely-used classification methods on BCI applications [18]. In particular, we have evaluated the LDA and SVM classifiers using the obtained features from each individual extraction method (PSD, Hjorth and AAR). Besides that, we consider an additional approach: LDA and SVM classification considering all the extracted features using the PSD, Hjorth and AAR methods.

In order to carry out fair comparisons with the proposed approach, we have used a LOO CV procedure for performance evaluation [13, 51], which is appropriate as the amount of available data is limited. It should be noted that our method makes use of the LOO CV technique for feature and model selection but, however, we cannot use the same LOO CV estimate for both feature/model selection and performance evaluation as this would introduce a (possibly quite strong) selection bias in favour of the proposed approach. Due to this, and as it is shown in Fig. 4, a nested LOO CV procedure is used, where an inner CV loop is used to perform the feature/model selection while an outer separate CV is used to compute an estimate of the classification error. Specifically, for the N total number of trials involved in the study, one is left out for testing, and the remaining $N-1$ are used for designing purposes. Under normal conditions, from these $N-1$ samples, $N-1$ different training subsets will formed by each time leaving one more sample out, giving us $N-2$ trials in each training subset. Therefore, the use of this nested LOO CV procedure in our proposed method would entail $N \times (N-1)$ iterations and it would be essentially impractical due to its excesive computational requirements. Nevertheless, as the experimental results show, the proposed approach is completely feasible for BCI applications thanks to the fact that the inner CV loop is automatically given by the Allen's PRESS statistic in a very fast and accurate manner. It is important to remark that the nested LOO CV strategy of the proposed approach ensures an almost unbiased estimate of the generalization performance of the proposed classification approach to new EEG signals.

All simulations have been carried out in MATLAB 7.11 (R2010b) environment running in the same machine with 4 GB of memory and 2.67 GHz processor. Note that the three feature extraction methods have been implemented using BioSig (v2.70),[2] an open source software library for biomedical signal processing. Moreover, the LDA (standard Fisher solution)

[2] http://biosig.sourceforge.net/

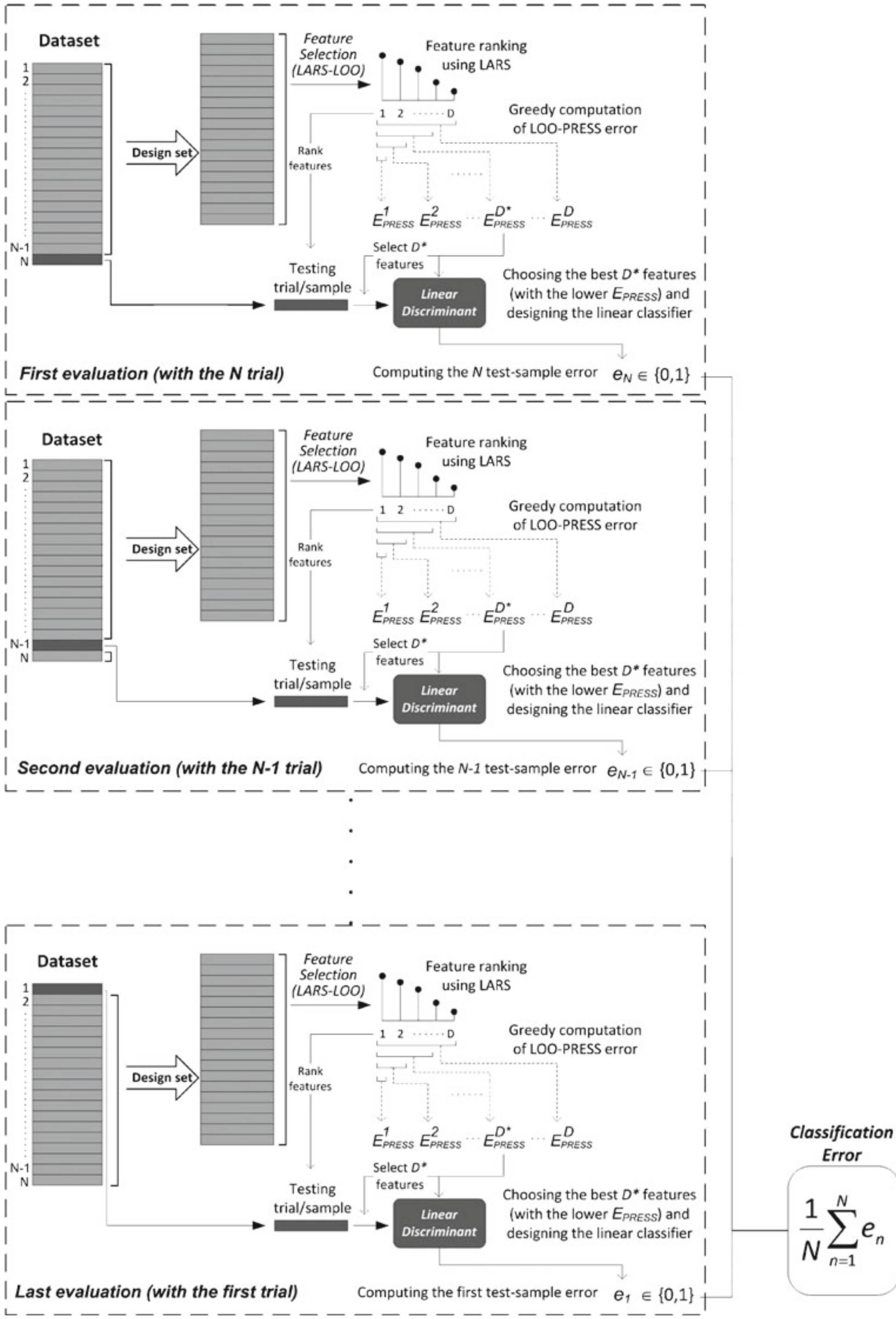

Fig. 4 Scheme of the procedure for performance evaluation of the proposed approach using a nested leave-one-out cross validation

and the SVM (with linear kernels and SMO optimization method) classification algorithms have been implemented using the MATLAB Bioinformatics Toolbox.

Experimental results

Table 1 shows the obtained LOO classification results (in %) by the proposed approach and the LDA/SVM classifiers trained using the extracted features with PSD, Hjorth and AAR techniques. The corresponding computational time (in seconds) is also shown in brackets. If the three feature extraction methods are individually analyzed, the PSD method provides better performance results with both LDA and SVM classifiers for all subjects, with the exception of the subject B07, where the Hjorth technique is most appropriate

Table 1 LOO classification accuracy results (in %) using LDA and SVM for the nine subjects by considering the tree different feature extraction procedures (PSD, Hjorth and AAR) and all the extracted features

Classification approach	Subjects B01	B02	B03	B04	B05	B06	B07	B08	B09	Mean
LDA+PSD	*76.2 (6.4e–1)*	59.2 (7.8e–1)	57.0 (5.0e–1)	91.4 (5.4e–1)	67.4 (6.7e–1)	69.7 (4.9e–1)	58.2 (5.1e–1)	*75.9 (5.6e–1)*	*70.0 (5.1e–1)*	69.4 (5.0e–1)
LDA+Hjorth	65.2 (5.3e–1)	56.0 (5.6e–1)	55.7 (5.3e–1)	73.3 (5.6e–1)	60.9 (5.7e–1)	52.5 (5.3e–1)	59.0 (5.3e–1)	67.3 (5.9e–1)	66.5 (5.3e–1)	61.8 (5.5e–1)
LDA+AAR	71.2 (6.3e–1)	51.0 (6.3e–1)	51.0 (6.3e–1)	79.0 (6.9e–1)	63.1 (7.0e–1)	57.8 (6.5e–1)	57.2 (6.3e–1)	68.7 (7.3e–1)	60.5 (6.3e–1)	62.2 (6.6e–1)
LDA+All Feat.	74.5 (8.0e–1)	*60.5 (8.3e–1)*	*57.5 (8.4e–1)*	*91.9 (9.0e–1)*	*70.5 (9.3e–1)*	*71.0 (8.4e–1)*	*60.7 (8.2e–1)*	74.3 (1.0e+0)	67.2 (8.5e–1)	*69.7 (8.7e–1)*
SVM+PSD	76.2 (1.3e+3)	58.5 (4.8e+2)	57.0 (2.5e+2)	*92.4 (1.1e+3)*	67.1 (2.5e+2)	69.5 (5.5e+2)	58.0 (2.5e+2)	*75.9 (1.8e+3)*	*70.7 (7.5e+2)*	69.4 (7.4e+2)
SVM+Hjorth	64.2 (1.2e+2)	55.8 (9.2e+1)	54.0 (6.9e+1)	79.3 (3.3e+3)	67.0 (7.1e+2)	53.0 (5.9e+2)	59.5 (7.0e+2)	66.8 (3.1e+2)	67.5 (3.4e+2)	63.0 (6.9e+2)
SVM+AAR	70.0 (5.1e+2)	51.0 (2.5e+2)	50.1 (2.5e+2)	79.1 (5.0e+3)	63.1 (1.1e+2)	57.5 (1.4e+2)	57.3 (2.5e+2)	68.5 (6.2e+2)	61.0 (1.6e+2)	61.9 (8.1e+2)
SVM+All Feat.	*77.7 (2.1e+3)*	*61.2 (1.5e+2)*	*57.3 (7.0e+2)*	90.9 (3.0e+3)	*71.7 (1.4e+3)*	*71.8 (1.5e+3)*	*61.8 (2.5e+2)*	75.0 (2.5e+3)	67.0 (1.3e+3)	*70.5 (1.4e+3)*
Prop. method	**79.8 (6.5e–1)**	**67.8 (5.8e–1)**	**62.0 (5.0e–1)**	**94.1 (5.9e–1)**	**76.2 (7.6e–1)**	**75.0 (6.1e–1)**	**65.8 (4.8e–1)**	**74.8 (5.4e–1)**	**72.8 (4.8e–2)**	**74.3 (5.8e–1)**

The last row corresponds to the obtained results of the proposed method and the last column shows the average results of each method. For each subject and LDA/SVM classifier, the best results from the three feature extraction methods are shown in italics. In each subject, best decision result is shown in bold face. Note that the total LOO computational time (in seconds) is also shown in brackets

for its LDA/SVM classifier modelling than the PSD and the AAR techniques. On the other hand, the combination of all feature extraction methods generally produces a slight performance enhancement on both LDA and SVM classification. As the average results of the nine subjects show (see the last column of Table 1), SVM exploits slightly better all the extracted features than the LDA classifier. Particularly, LDA achieves 69.4 % using the PSD features and 69.7 % using all features; whereas, the SVM classifier increases from 69.4 to 70.5 % using PSD and all extracted features, respectively. However, the design of a SVM classifier requires high computational capabilities. For example, under the LOO CV scheme, the average training time for LDA with all features is less than one second and, meanwhile, the SVM requires more than one thousand seconds with the same input dataset. Then, LDA is more appropriate than SVM as it generally provides good results with far fewer computational requirements.

According to the obtained results of the proposed method (see the last row of Table 1), it clearly outperforms the LDA and SVM approaches in terms of recognition rates as well as computational time. Except for the subject B08, the proposed method exploits better the combination of the three feature extraction techniques than LDA and SVM by selecting the most useful features for the recognition of each subject-specific EEG signals. As an example, for the subject B02, 60.5 % and 61.2 % are respectively achieved using LDA and SVM with all features; whereas, the proposed method achieves 67.8 %, which is an enhancement of 6.6 with respect to the SVM classification. From the average results of the nine subjects, the proposed method significantly improves the classification performance of the LDA and SVM approaches: 74.3 % vs. 69.7 and 70.5 %, respectively. It should be noted that this improvement does not entail a significant increasement of the computational requirements of the LDA approach. Proposed approach is very fast thanks to the efficient computational properties of the LARS algorithm and the Allen's PRESS statistic.

Finally, Fig. 5 shows the percentages of features (PSD, Hjorth and AAR) that have been selected for each subject using the proposed approach. The average results of the nine subjects are also shown. According to these results, in general, the PSD features are more useful than Hjorth and AR features. In some subjects, there is a significant reduction of the input feature space: more than 75 % of the input features are removed for the B03, B07, B08 and B09 subjects. Besides, the feature selection stage may entail that a feature extraction method is totally discarded for an specific

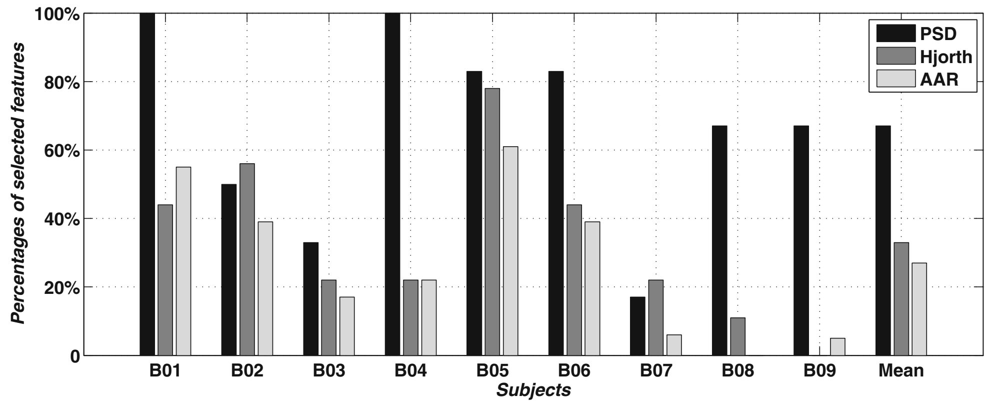

Fig. 5 Percentages of selected features from PSD, Hjorth and AAR methods using the proposed approach

subject, as in B08 and B09 where the AR and the Hjorth methods are respectively discarded. Therefore, from these obtained results, the proposed approach gives an appropriate subject-specific feature combination and selection which helps to provide better recognition rates and, then, a more accurate adaptive BCI system is achieved.

Conclusions and future works

This paper proposes an automatic and adaptive BCI system for classification of EEG signals. An accurate linear discriminant, with embedded feature selection, is presented and it is designed using a fast wrapper feature selection methodology based on the least angle regression (LARS) algorithm and the leave-one-out (LOO) cross validation procedure. The LARS algorithm gives an exact feature ranking for the linear classifier, which is used for selecting the most useful features for classifying EEG signals. The LOO error is computed by a direct and exact formula using the Allen's PRESS statistic and those features, which have been previously ranked using LARS, that minimize the LOO error are chosen for constructing the linear classifier. Three different widely-used feature extraction methods have been evaluated in parallel: PSD, Hjorth and AR. The proposed approach has been experimentally compared with the standard LDA and SVM classifiers using the BCI Competition IV dataset IIb. Obtained results shows the advantages, in terms of recognition rates and computational time, of our method over the state-of-the-art classifiers. The proposed approach allows to provide accurate and adaptive subject-specific BCI systems with low computational requirements thanks to the fast matrix computations of the LARS algorithm and the PRESS statistic. It should be noted that the design of our approach is automatic: there is not any tuneable parameter for the feature selection and classification stages.

As future work, it will be interesting to consider other feature ranking approaches, such as the Kullback–Leibler divergence or the Wilcoxon test. Moreover, all the experimental results of this research work have been obtained with an offline analysis and so they will have to be confirmed by online studies. Therefore, much of our future efforts will be aimed at the examination whether the proposed approach is suitable for online experiments with more channels and the problem of inter-subject variability.

References

1. Berger, H., Über das Elektroenkephalogramm des Menschen. *Arch. Psychiatr. Nervenkr.* 87:527–570, 1929.
2. Sanei, D., and Chambers, J., *EEG signal processing*. John Wiley & Sons, 2008.
3. Abibullaev, B., and An, J., Decision support algorithm for diagnosis of ADHD using electroencephalograms. *J. Med. Syst.* 1–14, 2012. doi:10.1007/s10916-011-9742-x.
4. Min W., and Luo G., Medical applications of EEG wave classification. *Chance* 22(4):14–20, 2009.
5. Tong, S., and Thankor, N. V., Qu*antitative EEG analysis methods and clinical applications*. Artech House, 2009.
6. Bashashati, A., Fatourechi, M., Ward, R. K., and Birch G. E., A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals. *J. Neural Eng.* 4(2):R32–R57, 2007.
7. Wolpaw, J. R., Brain-computer interfaces as new brain output pathways. *J. Physiol.* 579(3):613–619, 2007.
8. Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M., Brain-computer interfaces for communication and control. *Clin. Neurophysiol.* 113(6):767–791, 2002.
9. Carabalona, R., Castiglioni, P., and Gramatica, F., Brain-computer interfaces and neurorehabilitation. *Stud. Health Technol. Inform.* 145:160–176, 2009.
10. Machado, S., Araújo, F., Paes, F., Velasques, B., Cunha, M., Budde, H., Basile, L. F., Anghinah, R., Arias-Carrión, O., Cagy, M., Piedade, R., De Graaf, T. A., Sack, A. T., and Ribeiro, P., EEG-based brain-computer interfaces: An

overview of basic concepts and clinical applications in neurorehabilitation. *Rev. Neurosci.* 21(6):451–468, 2010.
11. VanErp, J. B. F., Lotte, F., and Tangermann, M., Brain-computer interfaces: Beyond medical applications. *Computer* 45:26–34, 2012.
12. Blankertz, B., Tangermann, M., Vidaurre, C., Fazli, S., Sannelli, C., Haufe, S., Maeder, C., Ramsey, L. E., Sturm, I., Curio, G., and Mueller, K. R., The Berlin brain-computer interface: Non-medical uses of BCI technology. *Front. Neurosci.* 4:1–17, 2010.
13. Duda, R. O., Hart, P. E., and Stork, D. G., *Pattern classification*. Wiley-Interscience, 2000.
14. Müller, K. R., Krauledat, M., Dornhege, G., Curio, G., and Blankertz, B, Machine learning techniques for brain-computer interfaces. *Biomed. Tech.* 49(1):11–22, 2004.
15. Cabrera, A., Farina, D., and Dremstrup, K., Comparison of feature selection and classification methods for a brain-computer interface driven by non-motor imagery. *Med. Biol. Eng. Comput.* 48:123–132, 2010.
16. Khorshidtalab, A., and Salami, M. J. E., EEG signal classification for real-time brain-computer interface applications: A review. In: *4th International Conference on Mechatronics (ICOM), Kuala Lumpur, Malaysia*. pp 1–7, 2011.
17. Lee, F., Scherer, R., Leeb, R., Neuper, C., Bischof, H., and Pfurtscheller, G., A comparative analysis of multi-class EEG classification for brain computer interface. In: *10th Computer Vision Winter Workshop*, 2005.
18. Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., and Arnaldi, B., A review of classification algorithms for EEG-based brain-computer interfaces. *J. Neural Eng.* 4(2):R1–R13, 2007.
19. Martens, S. M. M., and Leiva, J. M., A generative model approach for decoding in the visual event-related potential-based brain-computer interface speller. *J. Neural Eng.* 7(2):026003, 2010.
20. Krusienski, D. J., Sellers, E. W., Cabestaing, F., Bayoudh, S., McFarland, D. J., Vaughan, T. M., and Wolpaw, J. R., A comparison of classification techniques for the P300 speller. *J. Neural Eng.* 3(4):299–305, 2006.
21. Anderson, C., Stolz, E., and Shamsunder, S., Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. *IEEE Trans. Biomed. Eng.* 45(3):277–286, 1998.
22. Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., and Muller, K. R., Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Process. Mag.* 25(1):41–56, 2008.
23. Gandhi, T., Panigrahi, B. K., and Anand, S., A comparative study of wavelet families for EEG signal classification. *Neurocomputing* 74(17):3051–3057, 2011.
24. Hjorth B., EEG analysis based on time domain properties. *Electroencephalogr. Clin. Neurophysiol.* 29(3):306–310, 1970.
25. Zhao, H., Liu, C., Li, C., and Wang, H., Feature extraction using wavelet entropy and band powers in brain-computer interface. In: *2nd International Conference on Signal Processing Systems (ICSPS)*. Vol 2, pp V2-670–V2-673, 2010.
26. Schlögl, A., Vidaurre, C., and Müller, K. R., Adaptive methods in BCI research—an introductory tutorial. In: Allison, B., Graimann, B., and Pfurtscheller, G. (Eds.), *Brain-Computer Interfaces, The Frontiers Collection*. pp. 331–355, Springer, 2010.
27. Shenoy, P., Krauledat, M., Blankertz, B., Rao, R. P. N., and Müller, K. R., Towards adaptive classification for BCI. *J. Neural Eng.* 3(1):R13–R23, 2006.
28. Krusienski, D. J., Grosse-Wentrup, M., Galán, F., Coyle, D., Miller, K. J., Forney, E., and Anderson, C. W., Critical issue in state-of-the-art brain-computer interface signal processing. *J. Neural Eng.* 8:1–8, 2011.
29. Corralejo, R., Hornero, R., and Alvarez, D., Feature selection using a genetic algorithm in a motor imagery-based Brain Computer Interface. In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. pp 7703–7706, 2011.
30. Koprinska, I., Feature selection for brain-computer interfaces. In: *International Workshop on New Frontiers in Applied Data Mining (PAKDD), no. 5669 in LNCS*. pp 106–117, 2009.
31. Peterson, D. A., Knight, J. N., Kirby, M. J., Anderson, C. W., and Thaut, M. H., Feature selection and blind source separation in an EEG-based brain-computer interface. *EURASIP J. Appl. Signal Process.* 2005(19):3128–3140, 2005.
32. Sabeti, M., Boostani, R., Katebi, S., and Price, G., Selection of relevant features for EEG signal classification of schizophrenic patients. *Biomed Signal Proces* 2(2):122–134, 2007.
33. Schroder, M., Bogdan, M., Hinterberger, T., and Birbaumer, N. Automated EEG feature selection for brain computer interfaces. *In: First International IEEE EMBS Conference on Neural Engineering, Capri, Italy*. pp 626–629, 2003.
34. Efron, B., Hastie, T., Johnstone, L., and Tibshirani, R., Least angle regression. *Ann. Stat.* 32:407–499, 2004.
35. Bartoli, A., On computing the prediction sum of squares statistic in linear least squares problems with multiple parameter or measurement sets. *Int. J. Comput. Vis.* 85:133–142, 2009.
36. Tangermann, M., Müller, K. R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Miller, K., Müller-Putz, G., Nolte, G., Pfurtscheller, G., Preissl, H., Schalk, G., Schlögl, A., Vidaurre, C., Waldert, S., and Blankertz, B., Review of the BCI competition IV. *Front. Neurosci.* 6(55):1–29, 2012.
37. Ang K. K., Chin Z. Y., Zhang H., and Guan C., Mutual information-based selection of optimal spatial-temporal patterns for single-trial EEG-based BCIs. *Pattern Recognit.* 45:2137–2144, 2012.
38. Brodu, N., Lotte, F., and Lecuyer, A., Exploring two novel features for EEG-based brain-computer interfaces: Multifractal cumulants and predictive complexity. *Neurocomputing* 1: 1–12, 2011.
39. Pfurtscheller, G., Brunner, C., Schlögl, A., and da Silva, F. L., Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks. *NeuroImage* 31(1):153–159, 2006.
40. Oppenheim, A. V., and Schafer, R. W., *Discrete-Time Signal Processing, 2nd edn*. Prentice Hall, 1999.
41. Schlögl, A., *The electroencephalogram and the adaptive autoregressive model: Theory and applications*. Shaker Verlag, 2000.
42. Billinger, M., Brunner, C., and Neuper, C., Classification of adaptive autoregressive models at different sampling rates in a motor imagery-based BCI. In: *Fourth International BCI Meeting, Pacific Grove, CA, USA*, 2010.
43. Rodríguez-Bermúdez, G., Roca-González, J., Martínez-González, F., Peña Mora, L., Roca-González, J. L., and Roca-Dorda, J., Performance analysis of different feature-classifier binomials in motor-imagering BCIs: Preliminary results. In: *3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies, Rome, Italy*. pp. 1–5, 2010.
44. Delgado-Saa, J. F., and Cetin, M., Hidden conditional random fields for classification of imagery motor tasks from

EEG data. In: *19th European Signal Processing Conference (EUSIPCO 2011), Barcelona, Spain*. pp. 1377–1381, 2011.
45. Serre, D., *Matrices: Theory and Applications*. Springer, New York, 2002.
46. Blum, A. L., and Langley, P., Selection of relevant features and examples in machine learning. *Artif. Intell.* 97(1–2):245–271, 1997.
47. Fruitet, J., McFarland, D. J., and Wolpaw, J. R., A comparison of regression techniques for a two-dimensional sensorimotor rhythm-based brain-computer interface. *J. Neural Eng.* 7(1):16,003, 2010.
48. Garrett, D., Peterson, D., Anderson, C., and Thaut, M., Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* 11(2):141 –144, 2003.
49. Miche, Y., Sorjamaa, A., Bas, P., Simula, O., Jutten, C., and Lendasse, A., OP-ELM: Optimally pruned extreme learning machine. *IEEE Trans. Neural Netw.* 21(1):158–162, 2009.
50. Miche, Y., VanHeeswijk, M., Bas, P., Simula, O., and Lendasse, A.,TROP-ELM: A double-regularized ELM using LARS and Tikhonov regularization. *Neurocomputing* 74(16):2413–2421, 2011.
51. Lemm, S., Blankertz, B., Dickhaus, T., and Müller K. R., Introduction to machine learning for brain imaging. *NeuroImage* 56:387–399, 2011.