Efficient feature selection and linear discrimination of EEG signals

Germán Rodríguez-Bermúdez∗, Pedro J. García-Laencina∗, Joaquín Roca-González†, Joaquín Roca-Dorda∗

∗Centro Universitario de la Defensa de San Javier (University Centre of Defence at the Spanish Air Force Academy), MDE-UPCT.
†Universidad Politécnica de Cartagena, Industrial & Medical Electronics Research Group.
Campus Antiguo Hospital de Marina - Muralla del Mar, s/n, 30202 Cartagena, Murcia, Spain.

Abstract

Brain Computer Interface systems (BCIs) based on Electroencephalogram (EEG) signal processing allow to translate the subject’s brain activities into control commands for computer devices. This paper presents an efficient embedded approach for feature selection and linear discrimination of EEG signals. In the first stage, four well-known feature extraction methods are used: Power spectral features, Hjorth parameters, Autoregressive modelling and Wavelet transform. From all obtained features, the proposed method efficiently selects and combines the most useful features for classification with less computational requirements. Least Angle Regression (LARS) is used for properly ranking each feature and, then, an efficient Leave-One-Out (LOO) estimation based on the PRESS statistic is used to choose the most relevant features. Experimental results on motor-imagery BCIs problems are provided to illustrate the competitive performance of the proposed approach against other conventional methods.

Keywords: Brain Computer Interface (BCI), Electroencephalogram (EEG), Power Spectral Density (PSD), Hjorth parameters, Autoregressive (AR) modelling, Continuous Wavelet Transform (CWT), Least Angle Regression (LARS), Leave-One-Out (LOO), PRESS statistic, Fisher Linear Discriminant (FLD).

1. Introduction

Brain Computer Interface systems (BCIs) allow the identification of patterns of activation generated by the users’ brain [1]. In general, BCIs have different steps: the acquisition of multichannel EEG signals, the artifact detection and signal preprocessing, the feature extraction and the classification of the feature vectors using statistical machine learning techniques. Many research efforts have been focused on the two last stages: feature extraction and classification [1, 2].

Initially, many features are extracted from different methods and from channels covering all potentially useful brain areas. Then, typical BCI classification problems are defined by a huge number of input features. Nevertheless, many of them may not provide relevant information for the decision task and, thus, enlarge the complexity of the classifier design. Besides, BCIs have to tend to be adaptive systems for each subject and scenario in order to attain high performance and, thus, increase acceptance in users [3]. Furthermore, many additional factors may produce high variability of the EEG signals (due to fatigue, change of task involvement, etc). For all these reasons, the classifier and the subset of selected features require a periodic accurate fast adjustment to these changes in the EEG signals.

This work introduces a new fast embedded adaptive method for selecting and classifying feature input vectors from EEG signals. A regularized linear classifier for BCI problems is designed using a wrapper methodology based on Least Angle Regression (LARS) and Leave-One-Out (LOO) techniques, which has been previously introduced by Miche et al. for constructing extreme learning machines [4]. In particular, we exploit this successful methodology to implement a fast and accurate linear discriminant with embedded feature selection for EEG signal classification. LARS is used for properly ranking each feature and, then, an efficient LOO estimation based on the PRESS statistic is used to choose the most relevant features [6]. 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 Allen’s PRESS statistic [6]. Then, this approach performs both feature selection and classification efficiently and automatically, which makes its implementation in BCIs easier. Another relevant contribution is the combination of the following representative feature extraction methods for motor imagery [1]: Power Spectral Density (PSD), Hjorth parameters, Autoregressive (AR) modelling and Continuous Wavelet Transform (CWT) coefficients; because most previous methods do not usually work with more than two of these feature extraction methods in parallel [1].

It should be emphasized that, although the regularized Least Square (LS) method with $L_1$ penalty term has been previously analyzed in other BCI research works [7, 8, 9], this letter introduces a new efficient, automatic and fast wrapper-based feature selection framework for $L_1$ regularized linear discrimination of EEG signals, which is based on the combined use of the LARS algorithm and the Allen’s formula for the LOO error. To the

∗Corresponding Author.
Email addresses: german.rodriguez@cud.upct.es (Germán Rodríguez-Bermúdez), pedroj.garcia@cud.upct.es (Pedro J. García-Laencina)
best knowledge of the authors, this framework has not previously proposed for implementing BCI systems.

2. Feature extraction from EEG signals

The first stages of signal acquisition, artifact detection and preprocessing are omitted in this paper for the sake of clarity as they do not differ from any other BCI systems. This section introduces the basic notions of four representative feature extraction methods that, according to Bashashati et al. [1], have been widely applied in many BCI systems for motor imagery tasks.

- **Power Spectral Density (PSD) features.** From the four bands of an EEG signal, the most discriminative information for motor imagery is concentrated on the energy of the α (8-13 Hz) and β (13-30 Hz) bands [2]. The first step is to compute the Fast Fourier Transform (FFT) of the EEG signals. Then, the corresponding coefficients of the α and β bands are added.

- **Hjorth parameters.** In 1970, Hjorth introduced a set of three parameters to describe the EEG signal on the time domain [10]: Activity is the signal power (which is wide band filtered); Mobility is the mean frequency; and Complexity is the change in frequency.

- **Autoregressive (AR) coefficients.** AR model is a powerful tool used for signal modelling, where each input sample is predicted by a weighted linear combination of the previous \( m \) samples. After consulting several references and taking into account our previous experimental works [11], model order (\( 1 \)) has been set to 6. It is important to remark that the use of different model orders does not entail significant improvements [12].

- **Continuous Wavelet Transform (CWT) coefficients.** CWT provides a good way to visualize and decompose EEG signals into measurable component events [13]. The extracted wavelet coefficients provide a compact representation that shows the energy distribution of the EEG signal in time and frequency. Multiplying each coefficient by the appropriately scaled and shifted wavelet yields the constituent wavelets of the original signal. Standard ‘Daubechies D2’ wavelet type has been used.

2.1. Feature combination and selection

Finally, it is important to remark that the number of features extracted from each individual channel is usually too large and, besides, depending on the problem to be solved, the number of channels recorded by a BCI system may be large too. In an experiment with 32 channels and features derived from the four above mentioned feature extraction procedures (for instance, 17 features from each channel), the total number of features, i.e., the input data dimensionality, is 544. This is a high number of input features and it is likely that some of them are redundant or even irrelevant with respect to the classification task [14]. Moreover, BCIs have to tend to be adaptive systems [3] and, then, a fast and accurate feature selection has to be performed for reaching this goal.

In last years, some adaptive BCI systems have been proposed by exploiting feature selection techniques, which can be grouped in two categories: filter [12, 14] and wrapper [15]. On the one hand, 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 they 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 feasible in some BCI applications [14] and it has been usually performed using genetic algorithms (GA) [15]. Nevertheless, the main disadvantages of using GA for feature selection are its excessive running time to produce accurate results and the fact that each run of the GA creates a different subset of features.

Next, in order to solve all these drawbacks, Section 3 introduces a new wrapper-based approach for efficient feature selection and linear discrimination of EEG signals.

3. Proposed method

Assume \( N \) labeled input vectors obtained from the set of EEG signals under study, \( \mathbf{S} = \{(\mathbf{x}_n, t_n)\}_{t=1}^N \), where \( \mathbf{x}_n = [x_{0n}, x_{1n}, \ldots, x_{Dn}]^T \) is the \( n \)-th pattern, a \( D + 1 \) column vector with components \( x_{0n} = 1 \) (which is the bias unit) and \( x_{dn} \) a certain feature value (with \( d = 1, \ldots, D \)); and \( t_n \) is the target label for \( \mathbf{x}_n \), considering two possible classes: \( C_1 \) and \( C_2 \). In order to exploit the feature diversity, the \( D \) input attributes can include continuous variables obtained from several feature extraction procedures in different channels of the BCI system. Most gain from a combination of different features is expected when the single features provide complementary information for the classification task. This work is mainly focused on linear discriminants, being the Fisher Linear Discriminant (FLD) the most popular classifier in BCIs and it has been found to produce very good and robust performance [16]. Before describing the proposed method, linear discrimination is briefly introduced.

3.1. Notions of linear discrimination

The well-known representation of a linear discriminant is \( y_n = \mathbf{w}^T \mathbf{x}_n \), where \( \mathbf{w} = [w_0, w_1, \ldots, w_D] \in \mathbb{R}^{D+1} \) is known as weight vector. An input vector \( \mathbf{x}_n \) is assigned to class \( C_1 \) if \( y_n \geq 0 \) and to class \( C_2 \) otherwise. According to [17], the FLD can be also obtained by the linear least squares method if the target labels, \( t_n \), for class \( C_1 \) are set to \( N/N_1 \) and to \( -N/N_2 \) for \( C_2 \), where \( N_1 \) and \( N_2 \) are the numbers of samples of classes \( C_1 \) and \( C_2 \), respectively. Then, \( \mathbf{w} \) for the FLD is given by [17]:

\[
\min_{\mathbf{w}} \sum_{n=1}^{N} (y_n - t_n)^2 = \min_{\mathbf{w}} \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}_n - t_n)^2. \tag{1}
\]
Its Ordinary Least Squares (OLS) solution is given by \( \hat{w} = (X^T X)^{-1} X^T t = X^\dagger t \), where \( X \) is the \( N \times (D + 1) \) input data matrix (note that the \( n \)-th row of this matrix is \( x_n \)), \( t \) is the target vector and \( X^\dagger \) is the Moore-Penrose generalized inverse matrix of \( X \), assuming that \( X \) is full rank. Note that the Singular Value Decomposition (SVD) of \( X \) is used to compute the pseudoinverse for ensuring numerical stability and faster computations.

3.2. Feature ranking using LARS

In BCI applications, linear classifiers are generally more robust than non-linear models [16], since they have only limited flexibility (less parameters to set up) and, thus, are less prone to overfitting. However, linear systems can fail in the presence of outliers or strong noise situations, which is quite common in EEG data. For solving this drawback and to obtain better generalization capabilities, this work considers a regularization approach based on the \( L_1 \) penalty term or LASSO (least absolute shrinkage and selection operator) [7, 8, 9]:

\[
\min_{A, w} \left\{ \sum_{n=1}^{N} (w^T x_n - t_n)^2 + \lambda \sum_{d=1}^{D} |w_d| \right\};
\]

where \( \lambda \) is the regularization parameter. Note that (2) is a quadratic programming problem with linear constraints.

For solving (2), this work uses LARS (Least Angle Regression) [5, 18], which is a greedy algorithm that features one by one (according to a Least Squares criteria). The order in which the input variables are added to the linear discriminant provides a ranking of their usefulness for predicting the target data [5]. Compared to other classical ranking methods based on correlation measures, LARS is better as it specifically sorts the variables based on how much of the residual they can explain, i.e., how much new information they bring. According to [5], LARS finds the solutions for (2) in the entire regularization path (all possible values of \( \lambda \)), being its computational cost at the same order as solving the standard OLS problem [5, 18] without regularization or the optimization with single \( \lambda \).

3.3. Feature selection based on PRESS statistic

The correct selection of relevant features from EEG signals can help to design a classifier with better generalization performance [14]. This work makes use of an efficient wrapper methodology based on the previous feature ranking provided by LARS, which is unique for linear models [4]. According to this rank, the greedy algorithm adds the best feature at each round and the main control issue is to decide when to stop the incremental search, i.e., when no more features are incorporated to the model. In machine learning and applied statistics, this is typically done by Cross Validation (CV) procedures [17]. 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 [3]. It causes changes in the input feature space of the EEG data and, then, it may produce shifts from training to validation/test data. For these reasons, we may have to use Leave-One-Out (LOO) procedure for CV 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. For solving this inconvenient, and given that the model is linear, it is well-known that the Allen’s PRESS statistic [6] is a non-iterative and exact formula for computing the LOO error:

\[
\epsilon_{PRESS} = \frac{1}{N} \sum_{n=1}^{N} \left( \frac{t_n - x_n(x^T X)^{-1} x_n^T f_n}{1 - x_n(x^T X)^{-1} x_n^T} \right)^2;
\]

which can be implemented in an efficient way using:

\[
C = (x^T x)^{-1} \quad P = XC; \quad w = CX^\dagger t \quad D = \text{diag}(PX^T); \quad \epsilon = \frac{1}{N} x_n w, \quad \epsilon_{PRESS} = \frac{1}{N} \sum_{n=1}^{N} \epsilon_n^2.
\]

In this work, (3) is iteratively computed by adding a feature (which is previously ranked in \( X \)) to the model. The model with \( D \) features obtains the lowest PRESS statistics and this model is considered optimal, i.e., \( \epsilon_{PRESS} < \epsilon_{PRESS} \forall d \in (1, 2, \ldots, D) \) and \( d \) is not equal to \( D \).

Finally, once the most useful EEG features have been chosen, linear discrimination is performed using this reduced subset of \( D \) features. Figure 1 depicts a scheme of the proposed method.

4. Experiments

The proposed method has been evaluated experimentally in four well-known BCI datasets: one from the BCI Competition II [19] and the remaining three from BCI Competition III [20] (see Table 1). In each dataset, we have worked with the labelled EEG signals, from two bipolar channels C3 and C4, corresponding to different motor imagery states with feedback.

<table>
<thead>
<tr>
<th>Subject</th>
<th>BCI Comp. (set)</th>
<th>Trials</th>
<th>Classes</th>
<th>Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>II (III)</td>
<td>140</td>
<td>2</td>
<td>Bar</td>
</tr>
<tr>
<td>II</td>
<td>III (IIIb)</td>
<td>540</td>
<td>2</td>
<td>Basket</td>
</tr>
<tr>
<td>III</td>
<td>III (IIIb)</td>
<td>540</td>
<td>2</td>
<td>Basket</td>
</tr>
<tr>
<td>IV</td>
<td>III (IIIb)</td>
<td>320</td>
<td>2</td>
<td>Bar</td>
</tr>
</tbody>
</table>

Table 1: Datasets.
Table 2: LOO classification accuracy results (in %) using FLD and SVM for the four subjects by considering different feature extraction procedures (first four rows). For each subject and classifier, the best results from the four feature extraction methods are shown in italics. The last three rows correspond to the majority vote scheme of the four classifiers, the FLD/SVM classifier trained with the reduced data using PCA and the FLD/SVM classifier trained with all input features. 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.

<table>
<thead>
<tr>
<th>Feature extraction</th>
<th>Subject I</th>
<th>Subject II</th>
<th>Subject III</th>
<th>Subject IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FLD</td>
<td>SVM</td>
<td>FLD</td>
<td>SVM</td>
</tr>
<tr>
<td>PSD</td>
<td>73.57 (7.5e–2)</td>
<td>71.43 (2.4e+0)</td>
<td>55.94 (1.9e–2)</td>
<td>55.62 (5.9e+0)</td>
</tr>
<tr>
<td>Hjorth</td>
<td>65.71 (7.9e–2)</td>
<td>65.00 (8.9e–1)</td>
<td>57.81 (2.1e+1)</td>
<td>57.81 (1.0e+1)</td>
</tr>
<tr>
<td>AR</td>
<td>72.86 (8.8e–2)</td>
<td>73.57 (7.1e+0)</td>
<td>59.69 (2.2e+1)</td>
<td>58.75 (1.2e+1)</td>
</tr>
<tr>
<td>CWT</td>
<td>71.42 (8.5e–2)</td>
<td>69.29 (4.1e+0)</td>
<td>54.69 (2.4e+1)</td>
<td>55.93 (8.1e+0)</td>
</tr>
</tbody>
</table>

Majority Vote
- Subject I: 75.00 (3.2e–1)
- Subject II: 73.57 (1.5e+2)
- Subject III: 58.44 (8.7e–1)
- Subject IV: 66.85 (8.9e–1)

All Features
- Subject I: 80.71 (1.2e+0)
- Subject II: 75.74 (5.4e+1)
- Subject III: 61.88 (8.5e–1)
- Subject IV: 76.12 (3.6e+3)

Table 3: Obtained LOO classification accuracy results (in %) using the proposed method. Total computational time (in seconds) is also shown in brackets.

<table>
<thead>
<tr>
<th></th>
<th>Subject I</th>
<th>Subject II</th>
<th>Subject III</th>
<th>Subject IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FLD</td>
<td>SVM</td>
<td>FLD</td>
<td>SVM</td>
</tr>
<tr>
<td>Subject I</td>
<td>82.14 (8.9e–2)</td>
<td>67.18 (2.7e–1)</td>
<td>77.78 (7.5e–1)</td>
<td>68.95 (7.4e–1)</td>
</tr>
</tbody>
</table>

4.1. Experimental results

The experiments have been realized in order to show that the proposed approach is generally feasible for linear discrimination of EEG signals and, also, its advantages (in terms of recognition rates and required computational time) over other widely-used approaches to design BCIs for motor imagery.

For all subjects and EEG data from the datasets of Table 1, this work uses the standard analysis of the 1-s window during the motor imagery period, and the standard interval values for the α and β frequency bands. Then, after EEG signals are processed, 34 input attributes are extracted for each trial using the four feature extraction procedures described in Section 2. In particular, the input feature vector of each trial is composed of 4 features associated to PSD, 6 features corresponding to Hjorth parameters, 12 features obtained from AR modelling and the remaining 12 features computed from Wavelet analysis. Note that each trial is composed of two EEG signals, which are respectively associated to C3 and C4 channels, and, then, each feature extraction procedure is applied to both signals. Table 2 shows the LOO classification accuracy results (in %) obtained by conventional methods and its corresponding computational time. Initially, we have evaluated the classification accuracy obtained by FLD using each individual subset of features extracted from each method. Besides that, and in order to compare with these FLD classifiers, we have also considered a Support Vector Machine (SVM) for performing the classification stage given the same datasets. Then, classification is performed using the majority vote scheme with the four resulting FLD/SVM classifiers and, also, the FLD/SVM classifier considering all input features and the reduced dataset using Principal Component Analysis (PCA) [17]. As it has been already mentioned, classification accuracy has been computed using the LOO-CV procedure. 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 PSD, Hjorth and AR features have been extracted using the BioSig Toolbox¹, which is an open-source reference software for biomedical signal processing, and the CWT features using the MATLAB Wavelet Toolbox. FLD and SVM classifiers have been implemented using the MATLAB Bioinformatics Toolbox. In the case of the FLD, the classical Fisher solution has been used. With respect to the SVM classifier, gaussian kernels are chosen with a default scaling factor equal to 1.0 and the SMO method is used to train the SVM by considering 1.0e – 3 for the tolerance value with which the Karush-Kuhn-Tucker (KKT) conditions are checked and 15000 for the maximum number of the iterations.

According to Table 2, there is not a feature extraction method that is most appropriate regardless of the classifier and the subject. For subject I, FLD and SVM work better with PSD and AR features, respectively. AR features are the most suitable for both classifiers with the subjects II and III. For subject IV, CWT provides better results with FLD and SVM. From Table 2, a majority vote scheme does not ensure a recognition rate improvement of the individual classifiers performance. For example, when the majority vote scheme is used in subject IV, its performance is worse than the FLD+CWT and SVM+CWT solutions because classifiers based on PSD and Hjorth features provides very inaccurate results which degrades the classification performance of the voting scheme. With respect to feature reduction using PCA, it helps to provide better classification results with FLD and SVM. Nevertheless, the use of all inputs for designing the classifiers improves the recognition rates in most subjects, except in subject II where the SVM+PCA provides a slightly better result. FLD is more appropriate than SVM as it generally gives very good results (or better) with far fewer computational requirements. Only for subject II, SVM gives significantly higher classification results than FLD.

Table 3 gives the obtained LOO classification results using the proposed method and its corresponding computational time. According to these results, this method clearly outperforms the previous approaches in terms of recognition rates as well as computational time. This improvement has been achieved in

---

¹http://biosig.sourceforge.net/
less than one second in the four subjects. This method efficiently selects the most suitable features for linear discrimination of EEG signals using the ranking provided by LARS and the LOO estimation given by the PRESS statistic. The final number of selected input features is 13 for subject I, 15 for subject II, 22 for subject III and 21 for subject IV. Figure 2 shows the percentage of features that have been selected from each feature extraction method in each subject. In general, PSD and AR features are more suitable than Hjorth and CWT.

![Figure 2: Selected features (in %) from each method in each subject.](image)

5. Conclusions

This work presents an efficient method, with reduced computational complexity, for feature selection and linear discrimination of EEG signals in BCI applications. Proposed approach is based on the exact feature ranking given by LARS and the direct LOO error estimation by the PRESS statistic. Different feature extraction methods have been applied: PSD, Hjorth, AR and CWT. A robust selection and combination of these feature extraction methods for linear classification models in off-line analysis has been found to improve the classification accuracy of single EEG trials in comparison to (1) linear and SVM classification from each individual feature extraction method, (2) linear and SVM classification from all feature extraction methods and the reduced dataset based on PCA, and (3) voting of different linear and SVM classifiers corresponding to each feature extraction method. Most of our current efforts are to examine whether the proposed approach is suitable for on-line experiments with more channels and the problem of inter-subject variability [21]. It should be emphasized that on-line variants of the machine learning approaches used in this letter already exist [22, 23, 24, 25, 26] and, then, the proposed methodology can be directly extended for on-line experiments with recursive methods [24, 25]. Another ongoing research works are to consider other feature ranking approaches, like Kullback-Leibler divergence or Mutual Information, and to extend the proposed approach for other EEG applications.

References