Biol Cybern (2012) 106:389–405
DOI 10.1007/s00422-012-0506-6

ORIGINAL PAPER

Biological Cybernetics

Regularized logistic regression and multiobjective variable selection for classifying MEG data

Roberto Santana · Concha Bielza · Pedro Larrañaga

Received: 1 February 2012 / Accepted: 25 June 2012 / Published online: 2 August 2012

Abstract This paper addresses the question of maximizing classifier accuracy for classifying task-related mental activity from Magnetoencelophalography (MEG) data. We propose the use of different sources of information and introduce an automatic channel selection procedure. To determine an informative set of channels, our approach combines a variety of machine learning algorithms: feature subset selection methods, classifiers based on regularized logistic regression, information fusion, and multiobjective optimization based on probabilistic modeling of the search space. The experimental results show that our proposal is able to improve classification accuracy compared to approaches whose classifiers use only one type of MEG information or for which the set of channels is fixed a priori.

Keywords Brain computer interface · MEG · Multiobjective optimization · Classification · Feature subset selection · Probabilistic modeling

1 Introduction

The practical and scientific implications of using brain electrical activity as a way to interact with the external world are numerous, and their investigation is still at an early stage. Brain computer interfaces (BCIs) (Lebedev and Nicolelis 2006; Wolpaw et al. 2002) can translate electrical signals into commands without the need for motor intervention. They have been used intensively to provide communication and control to people with severe muscular or neural handicaps (Hoffmann et al. 2008; Iturrate et al. 2009; Nicolelis 2003), but they can also be used, for example, to conduct cognitive experiments (Carmena et al. 2003; Tan et al. 2009), improve human behavior, and facilitate interaction in special environments (Rossini et al. 2009), etc.

For analysis, a BCI can be divided into a signal acquisition module and a signal processing module (Wolpaw et al. 2002). Signal acquisition is executed using electroencephalography (EEG), MEG, or other techniques for recording brain activity. In this paper, we focus on the analysis of MEG data. Due to its cost and technical requirements, MEG is of limited use in practical BCI implementations. However, it is essential for investigating brain activity that cannot be extracted from EEG signals. In the BCI signal processing module, features are first selected over the original signal. The selected features are then translated into device commands. We will focus on the feature selection step and, in particular, on the conception of accurate and robust classification strategies able to deal with MEG data.

A variety of classification algorithms have been used to analyze brain data in the context of BCI applications (Lotte et al. 2007). Nevertheless, these methods have mainly been applied to EEG data where the number of sensors is usually smaller than in MEG. Support vector machines (SVM) (Vapnik 2000) and linear discriminant analysis (LDA) (Mclachlan 1992) are the two classifiers mostly applied to classification of MEG data.

Besserve et al. (2007) use a linear SVM classifier based on spectral power and synchrony features extracted from

R. Santana (✉)
Intelligent Systems Group, University of the Basque Country (UPV/EHU), P. Manuel de Lardizabal 1, 20018 San Sebastian, Spain
e-mail: roberto.santana@ehu.es

C. Bielza · P. Larrañaga
Computational Intelligence Group, Departamento de Inteligencia Artificial Universidad Politécnica de Madrid, Campus de Montegancedo, 28660 Boadilla del Monte, Madrid, Spain
e-mail: mcbielza@fi.upm.es

P. Larrañaga
e-mail: pedro.larranaga@fi.upm.es

continuous epochs of MEG data. SVM was also applied in Asano et al. (2009) to features extracted using an adaptive spatial filter approach. The MEG observations related to hand movement were initially prewhitened by the application of generalized eigenvalue decomposition that eliminated stationary interferences. Rieger et al. (2008) applied linear SVM to time- and wavelet-derived frequency representations of MEG data. The task was to predict, from single-trial-event-related magnetic fields recorded during the encoding of briefly visible natural scene photographs, whether a person would be able recognize the photograph later on.

Waldert et al. (2007) applied a regularized LDA to decode directions from MEG signals of the human contralateral motor cortex during center-out movements (four targets). Time domain features extracted from different time windows were used as inputs to the regularized LDA classifier. Bianchi et al. (2010) recently used MEG to investigate the evoked response components most suitable for use in a classical P300-based BCI interface speller protocol. They used a stepwise LDA fed with data relative to the first 800 ms of the signal following the visual stimulations. Wang et al. (2010) performed dimension reduction and MEG data transformation using an LDA that maximized linear discrimination among different movement directions.

The analysis of brain signals is frequently based on a priori knowledge about the physiological mechanisms that determine the brain activity (Wolpaw et al. 2002). Slow cortical potentials, P300, μ and β rhythms, and other types of electrophysiological signals used for BCI are associated with specific brain areas, and this information is implicitly or explicitly used by the signal acquisition or signal processing modules. However, there may be cases where the experimenter is interested not only in maximizing the accuracy of the mental task prediction based on the brainwave recorded data but also in investigating how information from different brain areas contributes to the predictions. Even if information is recorded from the same areas, subject and trial variability is frequently a source of poor BCI performance. In such cases, the interpretability of the machine learning techniques used for processing the data is even more essential.

This paper analyzes a machine learning technique that does not consider a priori information of how the brain data are related to the task under consideration. We address a classification problem whose objective is to predict, based on MEG data, the direction in which a subject is covertly focusing his or her attention. In this type of problem, attention is paid to a given stimulus without eye or head movement. It has recently been shown that high-accuracy classification, with potential BCI applications, can be achieved based solely on covert attention (van-Gerven et al. 2009; van-Gerven and Jensen 2009).

Our approach incorporates a number of novel alternatives for dealing with common problems experienced by BCI classification algorithms. To select a convenient (informative) original signal transformation procedure, we evaluate different ways to process the original signals (e.g., raw data, channel time series correlations, interaction graphs). To deal with noisy features and outliers and to increase the classifier generalization capabilities, we use a fast regularization-based classifier (Zou and Hastie 2005) that can deal with thousands of features. To further improve the classifier accuracy while trying to enhance robustness (accuracy variation across different subjects), we use a feature subset selection (FSS) method based on multiobjective optimization with probabilistic modeling of the search space. Finally, we propose new ways to extract physiologically relevant information from the learned classifiers.

Our analysis is in response to a challenge recently posed as part of a brain MEG data analysis competition.[1] Although the general approach described in this paper can be applied to other problems, we use the competition's task-related mental activity classification problem as an illustrative example of a successful application.

The paper is organized as follows. In Sect. 2, the general problem is described, and the experiments in which the brain data were acquired are explained. Section 3 introduces the main components of the proposed classification algorithm. The optimization approach to channel subset selection is explained in Sect. 4. The experimental framework, numerical results and discussion of the experiments are presented in Sect. 5, and work related to our proposal is briefly reviewed in Sect. 6. Finally, Sect. 7 concludes the paper.

2 Description of the problem

What follows is a general account of the experimental procedure. The data used in this paper were originally collected from the work presented in van-Gerven et al. (2009). For more details see van-Gerven et al. (2009).

2.1 Experimental framework

Fifteen subjects were instructed to covertly pay attention to different spatial locations of a screen (top, right, bottom, and left) during the registration of MEG information. The goal then was to analyze the recorded MEG data to detect, at the single-trial level, which of the four directions the subjects were paying attention to.

We focus on a 1-D version of the problem that was part of an open challenge to evaluate the accuracy of different

[1] The winner (ex aequo) of this challenge, held at the BIOMAG 2010 conference, was an implementation of the approach described in this paper. See http://megcommunity.org/index.php?option=com_content&view=article&id=2&Itemid=24 for details of the challenge.

machine learning techniques. The competition was focused on attention to the left and right (i.e., the problem was defined as one of binary classification), and data from only 4 of the original 15 subjects were used. The rules of the competition established that contestants should report the classification rate (proportion of correctly classified trials) for each subject as computed using leave-one-out cross validation and report the classification procedure. Artifact removal was allowed, but trials could not be rejected. Contestants were advised to prevent overfitting, e.g., if multiple algorithms were tried, then they were to be tested on the first subject and applied blindly to the remaining subjects. The data used for the challenge are described in van-Gerven et al. (2009).

Using the competition as a benchmark for introducing our approach, we are able to follow a clear evaluation methodology that is based on the competition rules and common to all the participants. It also serves to facilitate future comparisons with other methods since the experimental procedure and the data are publicly available.

2.2 Experimental data

The subjects viewed a screen with a central fixation cross and four squares at 7.5° of visual angle to the top, right, bottom, and left of the fixation cross. At regular intervals, a small arrow was displayed at the location of the fixation cross to indicate the direction to which subjects should covertly pay attention without moving their eyes from the fixation cross. A total of 128 trials were made per condition (top, right, bottom, and left) in eight subsequent sessions, interspersed by 1-min rests. Each trial started with a 400-ms presentation of the cue, after which subjects had 2,500 ms to covertly refocus their attention in the indicated direction. After this delay period, the square in the indicated direction turned either green or red for 40 ms. To facilitate task engagement and behaviorally measure task compliance, the subjects were asked to count the number of times the target location turned green over all eight sessions. There was a 1,500-ms rest between trials. The task was implemented in Presentation software (Neurobehavioral Systems, Albany, CA, USA).

Data were downsampled from 1,200 to 300 Hz. No further artifact rejection was performed. For each trial, the power spectrum was computed in the 5–70-Hz frequency range using a Hanning window for the period from 0.5 to 2.5 s after cue offset using 100-ms intervals. Preprocessed trials for the left and right conditions of each subject were available. Each trial was 2.5 s long and started −0.5 s before the cue, indicating which way the subject had to direct his or her attention. A total of 274 MEG channels were measured. Figure 1a shows a diagram of the location of the channels from which MEG information was extracted.

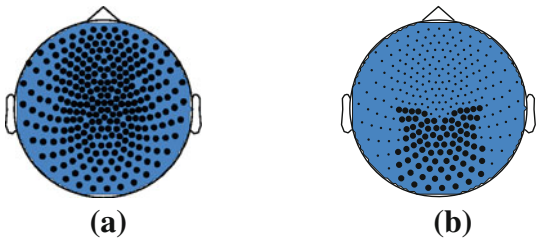

Fig. 1 MEG channel localization. **a** The complete set of 274 channels. **b** 86 Channels covering occipitoparietal brain areas

3 Factors in MEG data analysis

We distinguish three main factors that influence the classification accuracy:

- Type of information used for classification,
- Type of classifier,
- Channels from which the information is extracted.

In Sects. 3.1–3.3, these factors are explained, and we present the particular characteristics of our approach designed to take them into consideration.

3.1 Type of information used for classification

One of the elements that critically impacts classification is the particular information upon which the classification task is based. In our approach, we try different information processing variants before applying the classifier. In all cases, the starting point is the time series output from the $Nt = 128 \times 2 = 256$ trials, for $I = 274$ channels and $k = 4$ subjects. There are a total of $256 \times 274 \times 4 = 280,576$ time series. Each original time series comprises the period from −0.5 to 2.5 s at 100-ms intervals. Following van-Gerven et al. (2009), we use the period from 0.5 to 2.5 s following cue offset as the attention time only. This should counteract the influence of cue-evoked potentials. The MEG output data correspond to 600-component numerical vectors.

We apply four types of processing procedures (raw data, correlations between channels, interaction graphs constructed from correlations, and a representation that combines raw and correlation data) to the initial set of raw time series, i.e., each processing procedure tries to extract a different characteristic feature from the data.

3.1.1 Raw data

In this approach, the original set of 600 time points is reduced to a set of 60 components by averaging a time window comprising 10 points in each component. Following previous work Kelly et al. (2005) and van-Gerven and Jensen (2009)

where occipitoparietal alpha-band (8–14 Hz) EEG activity was used as a feature for left/right spatial attention classification, we assume that the relevant information for the classification is included in the range 0–14 Hz. Even after applying this modification we call the resulting information type "raw data." The classifier will receive a vector of $n = 60 \times 274 = 16,440$ features.

3.1.2 Correlations between channels

This approach takes advantage of any interaction between different brain regions during the solution of a recognition task.

We compute the correlation matrix between the time series corresponding to all the channels for a given trial. The correlation between two channels is computed as the correlation of their respective 600-component numerical vectors containing the channel measurements at each time point. A symmetric matrix $\mathbf{W}_{274\times274}$ is constructed for each trial. The classifier will receive a vector of $n = \frac{274\cdot273}{2} = 37,401$ features corresponding to the upper triangular part of the correlation matrix (without the main diagonal).

3.1.3 Interaction graphs constructed from correlations

The correlation matrix is used to construct interaction graphs between the different channels. The idea is that further analysis of the graph using topological measures from network theory can serve to reveal local and global information that is not directly recognizable from the correlation values.

The interaction graph $G = (V, A)$ is such that $V = \{v_1, \ldots, v_{274}\}$ is the set of vertices (channels) and the arc $a_{i,j} = (v_i, v_j)$ goes from vertex v_i to vertex v_j. Arcs are determined as follows:

$$a_{i,j} = \begin{cases} (v_i, v_j) & \text{if} \quad i < j \text{ and } \mathrm{cr}_{i,j} > 0.5 \\ (v_j, v_i) & \text{if} \quad i < j \text{ and } \mathrm{cr}_{i,j} < -0.5 \\ \text{no arc} & \text{otherwise} \end{cases}$$

where $\mathrm{cr}_{i,j}$ is the correlation coefficient between channels i and j.

The interaction graph is an arbitrary way to represent strong correlations (below −0.5 or above 0.5) between pairs of channels. We expect that if there are higher-order interaction patterns between the channels, at least some of them could be unveiled by a topological analysis of these graphs. These patterns could, in turn, be more informative for a classifier than raw data or pairwise correlations between the channels.

Figure 2a shows a possible correlation matrix for five channels. The corresponding interaction graph is shown in Fig. 2b.

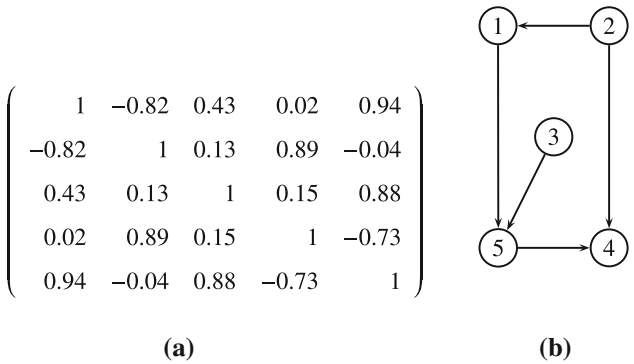

Fig. 2 Example of interaction graph construction. **a** Correlation matrix. **b** Interaction graph

Once interaction graphs have been constructed, the following local topological measures are computed for each node:

1. *Betweenness centrality* Measure of node centrality in graph. It is higher for vertices that occur on many shortest paths between other vertices.
2. *Pair distance* Average distance (defined as the length of the shortest path between two vertices) between each node and the other vertices. Disconnected vertices are assigned a very high, unattainable, distance value.
3. *Node eccentricity* Maximum of vertex finite distances to all other vertices.
4. *Clustering coefficient* Ratio of actually existing connections between the node's neighbors and the maximal number of such possible connections.
5. *Indegree* Mean indegree of vertices.
6. *Outdegree* Mean outdegree of vertices.
7. *Motif frequency, M=3* Motifs (Milo et al. 2002) are small network building blocks defined by their size M and interconnection patterns. We compute the motif frequencies of all motifs of size $M = 3$. Since only 5 of the 13 possible motifs appear at least once in all the graphs, these are the only motifs considered in our analysis. These motifs are shown in Fig. 3.
8. *Maximum modularity* Gives a modularity value corresponding to a network module decomposition computed using Newman's spectral optimization method, generalized to directed networks (Leicht and Newman 2008).
9. *Vertex participation coefficient* The participation coefficient (Guimera and Amaral 2005) defines how well distributed the links of a node are between different modules. It is close to 1 if the links are uniformly distributed among the modules and 0 if all the links fall within one module. The same modules used to compute the maximum modularity value are employed to compute the vertex participation coefficient.

In addition, a number of global topological measures are computed for the whole graph:

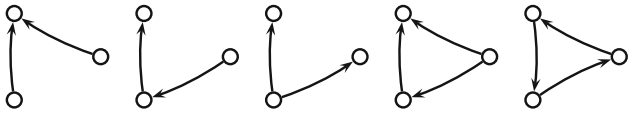

Fig. 3 All motifs ($M = 3$) that appear in interaction graphs learned from MEG data

1. *Assortativity coefficient* Computed as the Pearson correlation coefficient between pairs of linked nodes.
2. *Characteristic path length* Global mean of finite entries of graph distance matrix.
3. *Network radius* Minimum eccentricity of network vertices.
4. *Network diameter* Maximum eccentricity of network vertices.
5. *Network density* Average connection density of network, i.e., number of connections present in network out of all possible connections ($n^2 - n$).
6. *Number of vertices* Number of vertices in the network.[2]
7. *Number of edges* Number of edges in network.

The number of local features is $n_{\text{local}} = 274 \times 13 = 3,562$ and the number of global features is $n_{\text{global}} = 7$. The classifier receives $n = 3,569$ features, which is a considerably smaller number than in the previous two approaches. Computation of the topological measures is implemented using the brain connectivity toolbox (Sporns 2002). [3]

3.1.4 Approach based on raw and correlation information

We also try an approach conjointly using raw information and correlation coefficients between channels (Raw+Correlation representation). This implies the use of a vector of $n = 53,841$, which is a huge amount of features. This suggests the need to use efficient feature selection techniques to reduce the number of features.

3.2 Type of classifier

The response variable Z for the classification problem is binary (0 = subject is covertly paying attention to the left, 1 = subject is covertly paying attention to the right). The classifier of choice is a regularized logistic regression classifier in which the logistic regression sigmoid function represents the class-conditional probabilities through a linear function of the vector of predictor variables $\mathbf{v}$:

$$p(Z = 0|\mathbf{v}) = \frac{1}{1 + e^{-(\beta_0 + \mathbf{v}^T \boldsymbol{\beta})}}, \tag{1}$$

$$p(Z = 1|\mathbf{v}) = \frac{1}{1 + e^{+(\beta_0 + \mathbf{v}^T \boldsymbol{\beta})}} \tag{2}$$

$$= 1 - p(Z = 0|\mathbf{v}) \tag{3}$$

where β_0 is called the intercept and $\boldsymbol{\beta}$ the vector of regression coefficients. The model is fitted by regularized maximum (binomial) likelihood using an elastic net (Zou and Hastie 2005).

The elastic net solves the following general problem:

$$\min_{(\beta_0, \boldsymbol{\beta})} \left[\frac{1}{2N} \sum_{i=1}^{N} (z_i - \beta_0 - \mathbf{v}_i^T \boldsymbol{\beta})^2 + \lambda P_\alpha(\boldsymbol{\beta}) \right], \tag{4}$$

where N is the number of observations, $\lambda \in \mathbb{R}$, $0 \leq \lambda \leq 1$,

$$P_\alpha(\boldsymbol{\beta}) = (1 - \alpha)\frac{1}{2}||\boldsymbol{\beta}||_{l_2}^2 + \alpha||\boldsymbol{\beta}||_{l_1} \tag{5}$$

$$= \sum_{j=1}^{p} \left[\frac{1}{2}(1 - \alpha)\beta_j^2 + \alpha|\beta_j| \right] \tag{6}$$

is the elastic-net penalty, and p is the number of features. $P_\alpha(\boldsymbol{\beta})$ is a compromise between the ridge-regression or l_2 penalty ($\alpha = 0$) and the lasso or l_1 penalty ($\alpha = 1$).

3.2.1 Classifier evaluation and selection of lambda

To evaluate the classifier accuracy for a given data set, we use cross validation. Two possible alternatives are employed, leave-one-out and two-fold cross validation. We use leave-one-out cross validation as a thorough, definite validation of the classifier. Two-fold cross-validation is used as a faster *estimate* of the classifier accuracy. In this case, only one partition of the data set is used for the classification experiment. No variance of the classifier accuracy is computed. In Sect. 4.1 we explain the rationale behind the use of two-fold cross validation in this case.

An important issue for the elastic-net and other regularization techniques is the selection of the optimal α and λ values. In most of the experiments presented in this paper $\alpha = 1$, i.e., the lasso penalty is applied. Nevertheless, in Sect. 5, we present results on the influence of α on prediction accuracy. As regards the λ value, the elastic-net implementation we use outputs the prediction attained by the classifier for a set of λ values. This information can be employed to select the λ to be used to evaluate the test cases. In this paper we choose the λ that maximizes the accuracy of the training set. This means that the classification phase involves the computation of the model parameters and one additional validation step where these parameters are used to predict the outcome for the training set.

[2] Although this feature was automatically added to the classification vector and used in the experiments, we noticed later that it was not informative since all the graphs have the same number of vertices.

[3] http://sites.google.com/a/brain-connectivity-toolbox.net/bct/metrics.

3.3 Multiobjective FSS search

Channel selection will be used as a way to improve the classification results. On the one hand, we want to maximize the results accuracy for each of the subjects. On the other hand we would like to output a set of channels that, in terms of the results accuracy, is robust across individuals. Generally, intersubject variability determines that a subset of predictive features that works well for a given individual may produce poor results when used on a different subject. The set of optimal channels may also depend on the type of information selected.

To balance these two potentially conflicting goals, the optimal channels are searched using a multiobjective approach where each objective corresponds to the accuracy produced by the classifier for one subject. Each possible set of channels will have four (probably different) accuracy values, one for each subject. The question is then how to find a set of solutions $\mathbf{x}$ that can be considered accurate for at least some of the subjects and robust if all the subjects are considered. One possibility is to find the Pareto set of solutions, a common practice in multiobjective optimization (Coello et al. 2007).

Let a binary vector $\mathbf{x}$, with binary components $x_i \in \{0, 1\}$ $i \in \{1, \ldots, 274\}$, represent a possible selection of channels. $x_i = 1$ means that channel i has been selected to pass its corresponding information to the classifier, whereas $x_i = 0$ means that no information from channel i will be included in the classifier. We consider a maximization problem with $k = 4$ accuracy objective functions $f_i(\mathbf{x}) \rightarrow \mathbb{R}, i \in \{1, \ldots, k\}$, where the vector function $\mathbf{f}$ maps each solution $\mathbf{x} \in \mathcal{X} \in \{0, 1\}^n$ to an objective vector $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_k(\mathbf{x})) \in \mathbb{R}^k$.

In our application, each objective function $f_i(\mathbf{x})$ will correspond to the classification accuracy obtained for subject i when information extracted from the channels represented in $\mathbf{x}$ is used by the classifier. We expect that informative sets of channels will produce, on average, higher accuracies among all the subjects. However, it is also important to detect channels that are relevant for particular subjects. The Pareto set of solutions will contain the sets of channels that are globally and individually most informative. The concept of dominance is at the heart of the Pareto front approximation. It is assumed that the underlying dominance structure is given by the Pareto dominance relation "$\mathbf{y}$ dominates $\mathbf{x}$" that is defined as $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$, $\mathbf{x} \preceq_{\mathcal{F}} \mathbf{y} \iff f_i(\mathbf{x}) \leq f_i(\mathbf{y}) \, \forall i$, where $\mathcal{F} = \{f_1, \ldots, f_k\}$. The Pareto (optimal) set is given as $\{\mathbf{x} \in \{0, 1\}^n \mid \nexists \mathbf{y} \in \{0, 1\}^n \setminus \{\mathbf{x}\} : \mathbf{x} \preceq_{\mathcal{F}} \mathbf{y}\}$. It contains solutions that are nondominated. The associated Pareto front contains the vector of function evaluations for each of the Pareto set members. The extreme points of the Pareto set include the solutions that maximize each of the objectives. In our case, these solutions are the set of vectors that maximize accuracy for each of the subjects.

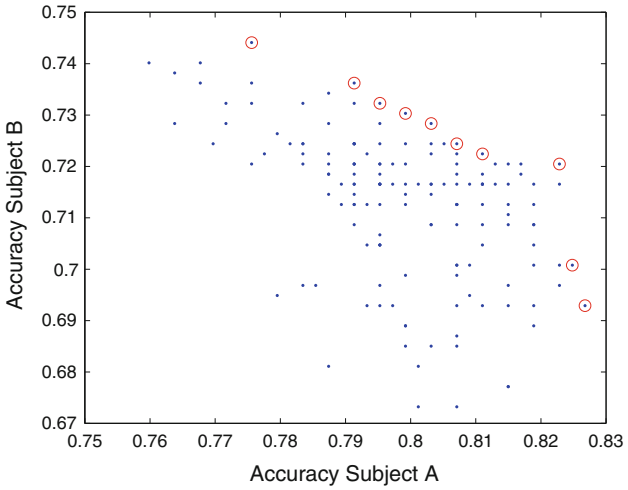

Fig. 4 Example of a Pareto front computed using as objectives the classification accuracies for two different subjects

The computation of the Pareto set is relatively simple. Each solution is compared to all other candidate solutions. If a solution is not dominated by any other solution from the candidate set, it belongs to the Pareto set of solutions; otherwise it is discarded from the Pareto set. Figure 4 shows an example of a Pareto front computed using the accuracies of only two subjects. In Fig. 4, each blue dot corresponds to a different set of channels. The location of the point is determined by the classification accuracies for two different subjects as computed using the information extracted from the respective channels. The 10 nondominated solutions that form the Pareto front are marked by a red circle.

We claim that the multiobjective approximation provides a wider perspective of the way in which intersubject variability operates. It also serves to identify individual features that consistently participate in solutions with a high accuracy. The optimization algorithms used to find a Pareto-set approximation are described in Sect. 4.2. Figure 5 shows how the classification of task-related mental activity for a single subject is accomplished using different sources of information. Figure 6 shows a diagram describing how each channel subset is evaluated using the classification accuracies computed for different subjects.

4 Channel subset selection

In the previous section we saw that accuracy results could be improved by appropriately selecting the brain regions from which channel information is fed to the classifier. In previous approaches (van-Gerven et al. 2009), this selection was made based on physiological knowledge about the brain areas thought to be involved in the mental task considered. We take a different approach to selecting the relevant channels.

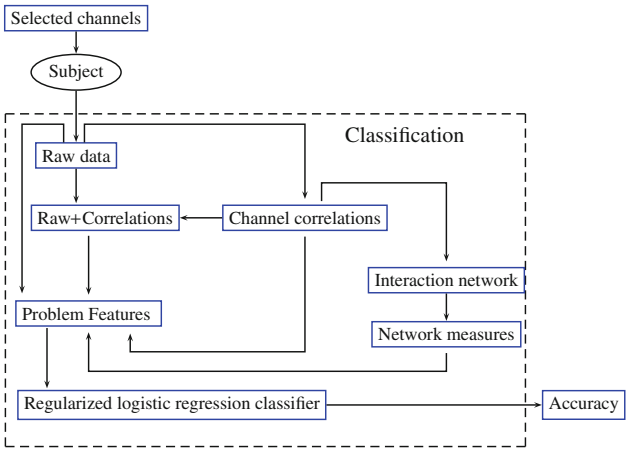

Fig. 5 Classification of task-related mental activity for a single subject using different sources of information

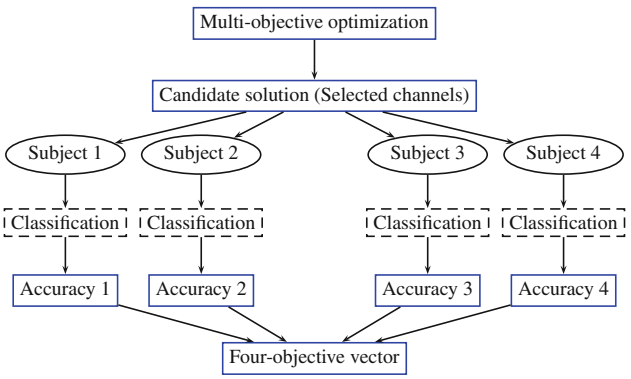

Fig. 6 Evaluation of candidate solutions in channel selection based on multiobjective optimization

Channel selection is posed as a multiobjective optimization problem where the feature multiset selection is carried out in a wrapper way. The quality of a candidate set of channels is based on the vector of four accuracy values, one accuracy value for each of the four subjects.

4.1 Problem representation and function evaluation

Since we intend to use a wrapper approach assisted by an optimization method (Saeys et al. 2007), two-fold cross validation is applied in place of the leaving-one-out cross-validation method. For the analyzed problems, leave-one-out is simply too costly to be affordable for an optimization heuristic, particularly for the large number of features considered in our case. Certainly, we can expect the same set of features to have different accuracy values when evaluated with two-fold or leave-one-out cross validation. However, we use two-fold cross validation as an estimate of the desired accuracy metric. This less accurate, but also less costly, metric will serve to guide the search for optimal solutions.

4.2 Genetic algorithms and estimation of distribution algorithms for multiobjective optimization

Evolutionary algorithms (EAs) are commonly applied to find Pareto-set approximations in multiobjective optimization problems. They use populations of solutions and apply selection based on the fitness of the solutions. We try three different EAs—one genetic algorithm (GA) (Goldberg 1989; Holland 1975) and two estimation of distribution algorithms (EDAs) (Larranhaga and Lozano 2002; Muhlenbein and Paaß 1996; Pelikan et al. 2002). GAs apply what are known as crossover and mutation operators to recombine solutions and visit new points from the search space. EDAs are similar to GAs. However, they replace traditional crossover and mutation operators by the estimation and sampling of probabilistic models. EDAs have been successfully applied to FSS problems (Armananzas et al. 2011; Inza et al. 2000; Mendiburu et al. 2006) and were recently proposed for application to problems of neuroscience (Santana et al. 2010a,b). The idea of using these three different optimization algorithms is that together they would allow us to try different ways of exploring the search space. For our analysis of the solutions, documented in the experimental section (Sect. 5), we took an equal number of executions from each algorithm and extracted the best solutions found from this complete set.

Our GA uses one-point crossover and bitwise mutation (Goldberg 1989). In the case of EDAs, the choice of the probabilistic model and the particular class of learning and sampling methods is fundamental. The models may differ in the order and number of the probabilistic dependencies that they represent. A variety of learning and sampling techniques can be used depending on the type of representation and other characteristics of the optimization problem. In particular, there may be important differences between EDA implementations for single and multiobjective problems. Enforcing the population diversity needed to guarantee a good covering of the Pareto set is particularly important for multiobjective problems, and specialized learning and sampling methods may be conceived to fulfill this goal.

Algorithms 1 and 2 respectively show the pseudocodes of GA and EDA for multiobjective optimization problems. In both algorithms, the selection method employed uses Pareto ranking selection (Coello et al. 2007) where individuals are ordered according to the Pareto front to which they belong. Individuals in the first front (nondominated solutions) come first, followed by individuals that are only dominated by others in the first front and so on. Within each front, they are ordered according to the average rank of their fitness functions. After the entire population has been ordered, truncation selection is applied to select the best T percentage of the population.

We use two different EDA variants. Each variant captures and uses different relationships between the problem

Algorithm 1: GA for multiobjective optimization

```
1  D_0 ← Sample M individuals using a uniform distribution
2  t ← 1
3  do {
4  Evaluate D_{t-1}
5      D^{Se}_{t-1} ← Select N individuals from D_{t-1} using Pareto ranking selection
6      Randomly select a mating-pool of individuals from the selected set
7      Generate D_t by applying recombination and crossover on the mating-pool
8  } until Stop criterion is met
```

variables, effectively implementing diverse search strategies. The first variant considered uses a univariate marginal product model in which all variables are independent, i.e., no dependencies are represented in the model. The joint probability distribution of the univariate marginal distribution algorithm (UMDA) (Muhlenbein and Paaß 1996) can be factorized as follows:

$$p_{\mathrm{UMDA}}(\mathbf{x}) = \prod_{i=1}^{n} p(x_i). \tag{7}$$

The second model learns a probabilistic model based on a tree. In this model, each variable may depend on no more than one variable, called the parent. The probability distribution $p_{\mathrm{Tree}}(\mathbf{x})$ used by Tree-EDA (Santana et al. 2001) is defined as

$$p_{\mathrm{Tree}}(\mathbf{x}) = \prod_{i=1}^{n} p(x_i \mid \mathrm{pa}(x_i)), \tag{8}$$

where $\mathrm{pa}(X_i)$ is the parent of variable X_i in the tree, and $p(x_i \mid \mathrm{pa}(x_i)) = p(x_i)$ when $\mathrm{pa}(X_i) = \emptyset$, i.e., when X_i is the root of the tree. Probabilistic trees can be represented by acyclic connected graphs.

Algorithm 2: EDA for multiobjective optimization

```
1  D_0 ← Sample M individuals using a uniform distribution
2  t ← 1
3  do {
4  Evaluate D_{t-1}
5      D^{Se}_{t-1} ← Select N individuals from D_{t-1} using Pareto-ranking selection
6      Learn a probabilistic model from D^{Se}_{t-1}
7      D_t ← Sample M individuals from the probabilistic model
8  } until Stop criterion is met
```

The stop criterion used for the algorithms was to reach a maximum number of generations.

The computational cost of multiobjective optimization EAs depends on the population size, the number of generations, and the evolutionary operators used. The main difference between the three variants of the used EAs is in the complexity of the algorithms used for combining the solutions during the reproduction step. The complexity of the GA crossover operator is linear in the selected population size, i.e., $O(N)$. The learning algorithm used by UMDA is linear in the selected population size and the number of variables, i.e., $O(Nn)$. Finally, the learning algorithm used by Tree-EDA is quadratic in the number of variables and linear in the selected population size, i.e., $O(Nn^2)$.

4.3 Extended approach: improving accuracy by augmenting the amount of information

In some cases, when a set of channels is given a priori or the channels have been found using a particular type of information, it would be interesting to find out how new different types of information added to the classifier would modify the classification. This will only be applied in situations where additional information is added to the classifier for which the currently used information is insufficient for achieving the targeted classification accuracy. Instead of searching the solution using the raw+correlation information, as discussed in Sect. 3.1.4, we search for the optimal set of channels using a particular type of information (as in Sects. 3.1.1–3.1.3). Once the optimal set of channels has been found, new features are added to the classifier. We apply this approach to find the set of channels using raw data, and once the optimal channels have been found, the classifier is invoked passing a vector comprising the original features selected from raw data and, *additionally*, the correlation features for each of the selected channels as features.

5 Experiments

In this section we investigate the combination of factors that produces the best classification accuracy both globally and for each of the subjects. Our analysis is focused on the type of information and the set of channels. In addition, we empirically investigate a number of issues that influence classification and should be taken into account to interpret the results produced by the algorithm. We start by presenting an overview of the experiments and the questions these experiments address. The subsequent sections present the results for these experiments and discuss the results.

5.1 Overview of the experiments

As stated in Sect. 3.2, the parameter α of the logistic classifier sets a compromise between the ridge-regression or l_2 penalty ($\alpha = 0$) and the lasso or l_1 penalty ($\alpha = 1$). A necessary initial step for the application of the classifier is evaluating the

influence of α in the accuracy results. In Sect. 5.2 we carry out this evaluation for all subjects and types of information.

A fundamental question we address is whether the use of different types of information can produce different results in terms of accuracy. This is an important step of the experimental procedure since one of the assumptions made in this paper is that the use of different sources of information can improve the classification results in the analysis of MEG data. A related question is whether the use of information from all the channels can be more effective in terms of accuracy than information constrained to a particular brain area (in this case, the occipitoparietal region). The questions of which is the most informative type of information and the right choice of the channels are very related. Therefore, we address them together in Sect. 5.3.

The next issue addressed in our experiments is whether the multiobjective approximation approach is able to detect the most informative channel sets for each of the subjects. To this end we run the three variants of the multiobjective EAs and compute the Pareto-set approximations using the combined output of these three variants. Then we compute the best accuracies achieved for each of the subjects and each type of information. This analysis is presented in Sect. 5.4, where we also inspect the Pareto sets and identify the best subsets of channels for each individual and type of information.

Another important question that we investigate is whether the channel sets contained in the Pareto sets found for a number of subjects can be useful in the classification of other subjects and with downsampled frequencies. We call this type of study a robustness analysis, and it is addressed in Sect. 5.5, where we evaluate the Pareto sets computed from the four subjects in a larger set of 15 subjects and using less information from the original brain signals.

Finally, in Sect. 5.6 we show how the parameters learned by the logistic classifier can be used for determining the different contribution of channels to the classification accuracy. Allowing one to determine the channel relevance is an added value of this type of classification algorithms. We further extend the analysis of the logistic classifier parameters by identifying, on the basis of these parameters, the time periods of the recorded time series that are more informative for the classification task.

All the optimization algorithms (GA, UMDA, and Tree-EDA) are implemented using the MATEDA-2.0 software (Santana et al. 2010c), a modular implementation of estimation of distribution algorithms programmed in Matlab (The MathWorks 2007) that can be used to implement genetic and other classes of EAs. The computation of all network measures is implemented using the brain connectivity toolbox (Sporns 2002). We use the Matlab implementation of the regularized logistic classifier proposed in Friedman et al. (2010), which uses cyclical coordinate descent, computed along the regularization path. Routines for data processing and analysis of the experiments were programmed by the authors in Matlab.

5.2 Study of the alpha parameter

As an initial step we investigate the effect of the parameter α on the accuracy of the classification results. For each subject and each type of information, classification accuracy is computed for $\alpha \in \{0.1, 0.2, \ldots, 1.0\}$. To reduce the computational time overheads, channel selection is constrained to the set of 86 occipitoparietal channels. Figure 7 shows the curves describing the variation in the accuracy as a function of α.

A first conclusion from analyzing Fig. 7 is that there are only minor differences in the accuracy values due to variations in α. We also find that raw+correlation information (black lines) achieves improvements in the classification rate over the correlation type of information (blue lines) for all the subjects. Compared with the use of raw information (red lines), the raw+correlation information clearly produces higher accuracy for subjects 1 and 2. However, for subjects 3 and 4, the difference is not so clear. We also observe that, at least in some situations, interaction graphs (green lines) can improve the classification given by correlation values. This applies to the fourth subject and may indicate that, in this case, higher-order patterns of correlations are captured by the computed graph measures. Based on these results, in the remaining experiments presented in this section, we arbitrarily set $\alpha = 1$, which corresponds to the lasso penalty.

5.3 Channels from which information is extracted

We conduct an exploratory set of experiments to evaluate the accuracy of the classifier for a changing number of channels and using the different types of information. We consider two different scenarios: (1) the classifier receives information from all channels (274) and (2) the classifier receives only information about the occipitoparietal channels (86) (Fig. 1b). Table 1 shows the results obtained for the accuracy where leave-one-out cross validation was always used.

Looking at Table 1, we find that the classifier accuracy is variable depending on the channels providing the information. Even if regularization implicitly makes a feature subset selection by setting to zero the coefficients of features that do not support relevant information for the classification task, channel preselection can improve the classification results.

In Table 1, we underline the cases where using a smaller number of channels (occipitoparietal channels) improves the accuracy of the classifier that uses all the information. Notice that when a channel is excluded, a complete set of variables (those features that represent the corresponding type of information extracted from that channel) is not given as classifier input. For instance, using raw data and constraining the

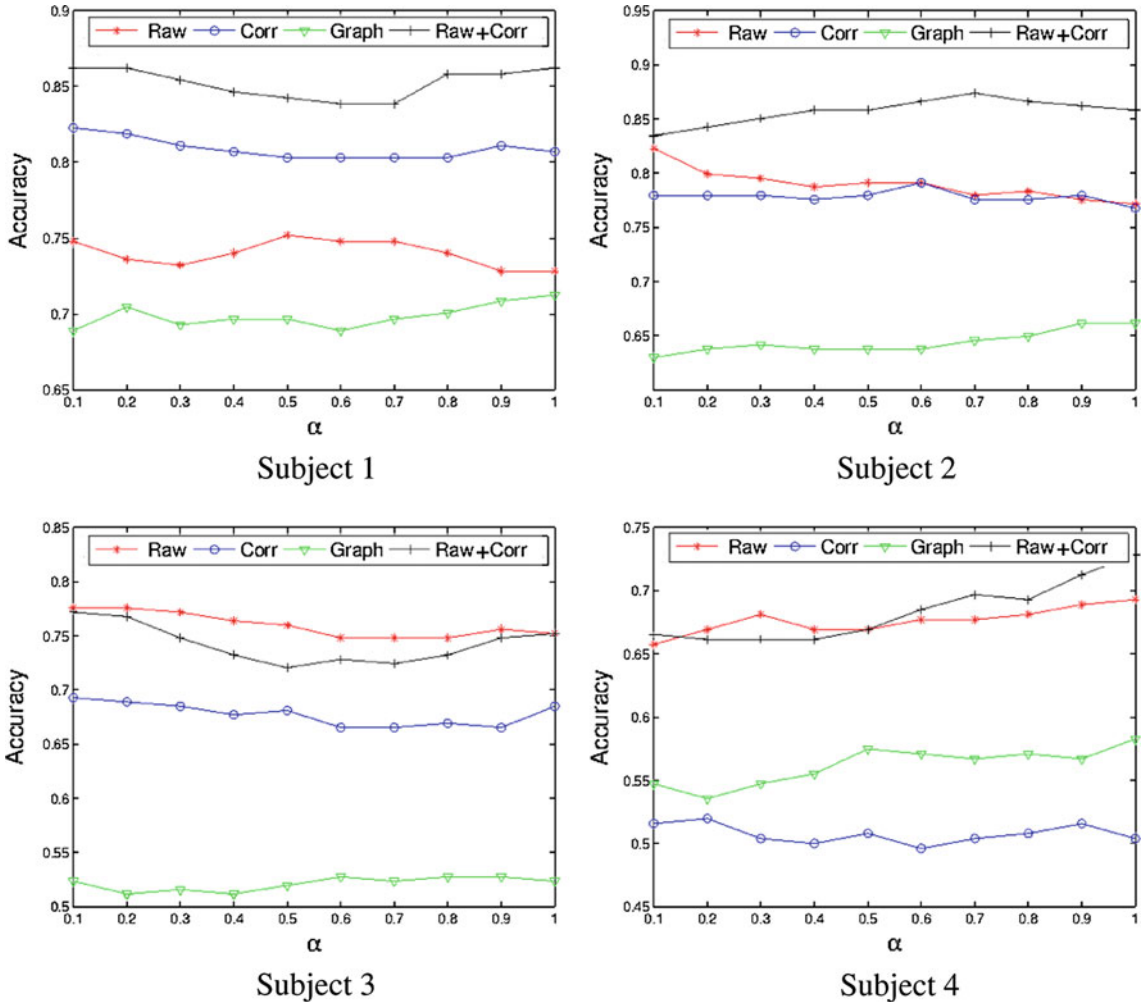

Fig. 7 Accuracy of the classification results as a function of parameter α for each of the subjects and each type of information. The initial channel set comprises the occipitoparietal channels

Table 1 Classification accuracy with all channels (All) and occipitoparietal channels (OP) for all the subjects (*rows*) and the four types of information used (in *columns*)

Subject	Raw		Correlation		Graph		Raw+Corr	
	All	OP	All	OP	All	OP	All	OP
1	0.7362	0.7283	0.8031	0.8081	0.6596	0.7126	0.8504	0.8622
2	0.7835	0.7717	0.8071	0.7677	0.5906	0.6614	0.8189	0.8583
3	0.7323	0.7520	0.6654	0.6850	0.4528	0.5236	0.7677	0.7520
4	0.7953	0.6929	0.5669	0.5039	0.4321	0.5827	0.7677	0.7283

number of channels to 86, this means that only $86 \times 60 = 5{,}160$ features will be fed to the classifier.

5.4 Classification experiments

In the next step, we apply multiobjective optimization to find a set of channels that, if fed to the classifier, outputs a high classification accuracy. For each subject and each type of information we run each of the three EAs 10 times. All EAs use a population size of $M = 50$ individuals, selection parameter $T = 0.5$, and a maximum of 100 generations. The solutions evaluated by these runs are the basis of a postprocessing step where the best channel sets are selected for each type of information. All the steps that lead to finding the best classifiers are described in Algorithm 3.

Notice that in step 4 the accuracies are computed using leave-one-out cross validation instead of the two-fold cross validation employed by the EAs. Consequently, PS_I may not

Algorithm 3: Steps for finding the best classifiers

1	Given a type of information l, run 10 executions of each multiobjective EA
2	Collect in the set A_l all the selected populations for each generation of the EAs
3	Find the Pareto set PS_l from A_l
4	Using leave-one-out cross-validation, estimate the set of classification accuracies $c_j(\mathbf{x})$, $j = 1, \ldots, 4$ for each vector $\mathbf{x} \in PS_l$
5	For each subject j, find the vector $\mathbf{x} \in \mathrm{PS}_l$ such that one $c_j(\mathbf{x})$ is maximized, $j = 1, \ldots, 4$
6	Find also the vector $\mathbf{x} \in \mathrm{PS}_l$ such that $\frac{1}{4}\sum_j c_j(\mathbf{x})$ is maximized

Table 2 Best classification results for the different types of information

Subject	1	2	3	4
Optimizing	Raw information			
1	0.7598	0.7992	0.7244	0.7441
2	0.6850	0.8425	0.7756	0.7874
3	0.7559	0.7677	0.7953	0.7677
4	0.6890	0.7913	0.7441	0.8425
Best mean	0.7244	0.7953	0.7835	0.8386
	Correlation values			
1	0.8425	0.7638	0.7087	0.6339
2	0.7953	0.8228	0.6890	0.5748
3	0.7953	0.7835	0.7520	0.5118
4	0.8189	0.7244	0.6811	0.6732
Best mean	0.8268	0.7795	0.7520	0.6417
	Interaction graphs			
1	0.7480	0.6535	0.5787	0.4646
2	0.6890	0.7126	0.5039	0.5315
3	0.6890	0.6339	0.6693	0.4882
4	0.6378	0.5945	0.5551	0.6063
Best mean	0.7087	0.6614	0.5827	0.5433
	Raw+Correlation information			
1	0.8701	0.8228	0.7638	0.7362
2	0.8031	0.8505	0.7480	0.7717
3	0.8228	0.8150	0.7913	0.7992
4	0.8228	0.8150	0.7913	0.7992
Best mean	0.8504	0.8150	0.7913	0.7874
	Extended information			
1	0.8701	0.7953	0.7480	0.7480
2	0.8661	0.8425	0.6693	0.7126
3	0.8386	0.8346	0.8031	0.7362
4	0.8228	0.8189	0.7323	0.8031
Best mean	0.8386	0.8346	0.8031	0.7362

be a Pareto set if the **c** vector values estimated in step 4 are taken as the objective values. The best accuracy value for each subject corresponds to extreme points of the Pareto front and are computed in step 5 of Algorithm 3. As a measure of global behavior, the average of the accuracies is computed in step 6.

Table 2 shows the objective vectors comprising the best accuracy for each of the subjects when each type of information is used. A row of the table displays the four accuracies estimated for a given channel set that is included in the Pareto set of solutions. For each type of information, five rows are presented. The first four rows respectively correspond to the solutions that maximize the accuracy for each of the four subjects. Therefore, the main diagonal of these four rows comprises the best accuracy achieved among all the channel sets for each subject. The last row (best mean) shows the objective vector with the highest average accuracy among the four subjects. Also, for each type of information, Table 3 summarizes this information by displaying the best absolute classification accuracies found for each subject. The best classification accuracy for each subject among all types of information is underlined. Results showed improvement over previously achieved accuracies reported in van-Gerven et al. (2009) for all the subjects. The improvement was particularly remarkable for subjects 3 and 4 for which previously best known accuracy values were around 0.72 and 0.65, respectively.

Table 3 Best absolute results achieved using all types of information

Subject	1	2	3	4	Type
1	0.8701	0.7953	0.7480	0.7480	Extended
1	0.8701	0.8228	0.7638	0.7362	Raw+Corr
2	0.8031	0.8505	0.7480	0.7717	Raw+Corr
3	0.8386	0.8346	0.8031	0.7362	Extended
4	0.8228	0.8189	0.7323	0.8031	Extended
Mean	0.8504	0.8150	0.7913	0.7874	Raw+Corrf

Figure 8 shows (as dot points) the scalp location of the best subset of channels learned for each individual using the raw+correlation information. The color represents the average accuracy of each channel computed from the Pareto front. The average accuracy of channel i is computed as the average accuracy of all objective vectors whose corresponding solution in the domain includes channel i. Green spots correspond to channels that were not present in any of the solutions comprised by the Pareto set. The figure indicates that green spots are equal across individuals, whereas the color of the other channels changes only slightly.

The information about the best single solution for each subject, as displayed in Fig. 8, is not very informative because the number of selected channels is relatively large and it is difficult to detect any particular pattern. One alternative for refining our analysis is to identify channels that are often in the solutions comprised by the Pareto set. Figure 9 shows (as dot points) the scalp locations of channels included in at least 80 % of the Pareto set for the four types of information considered. These are expected to be channels, for the considered type of data, that reliably provide relevant features for the classification task. In Fig. 9 the colors represent the frequency with which each channel is present in the Pareto sets.

Figure 9 shows that most of the channels frequently included in Pareto solutions are located around the occipito-parietal region. There are also channels selected from other

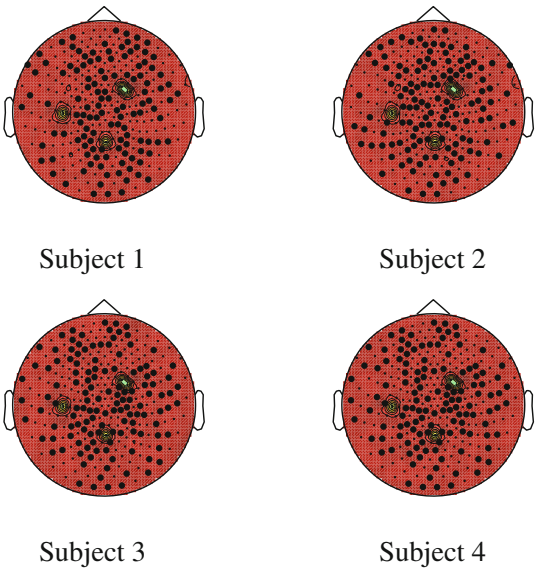

Fig. 8 Best subsets of channels learned for each individual using a raw+correlation information scheme

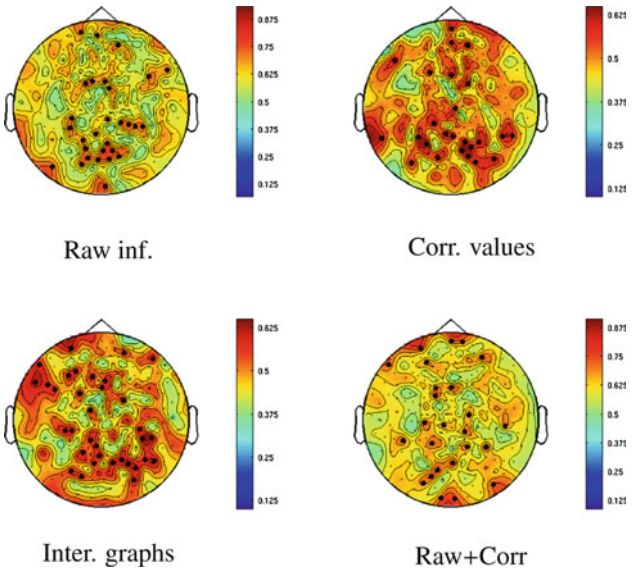

Fig. 9 Channels that were in at least 80 % of Pareto-set solutions for each information type

areas; in particular some channels are detected in the frontal area. These results seem to indicate that important gains in interpretability can be attained when a set of solutions is used instead of a single one. We must be aware, though, that the optimal solutions found by EAs and other multiobjective optimization methods are correlated by the way in which the search was conducted. This could mean that a channel is often in a Pareto set due to the way the EA works. This effect can be countered by using several runs or different variants of the search procedure, as in our approach.

5.5 Evaluating the robustness of the Pareto-set solutions

We are interested in further evaluating the robustness of the solutions included in the Pareto set. We focus the analysis on the set of 152 Pareto solutions found for the raw+correlation type of information and extend the evaluation of the classification accuracy to the original set of 15 subjects. For each of the 15 subjects, the brain signals are downsampled from 1, 200 to 60 Hz. This implies that we are using approximately one fifth of the signal information that was used in the previous classification experiments for which signals were downsampled from 1, 200 to 300 Hz. Therefore, we do not expect to achieve the same classification accuracy results. We do not expect neither that the Pareto sets of channels found for four subjects will necessarily be good for the other eleven subjects. However, we can compare them with the subset of occipitoparietal channels given the known physiological mechanism involved in covert attention.

Figure 10 shows the classification accuracy achieved in the complete set of 15 subjects and with the raw+correlation type of information for all the channels, the occipitoparietal channels, and the best solution of the Pareto set using only four subjects. The best solution depends on each new subject. It is the set of channels from the Pareto set that gives the highest accuracy (out of the 152 values) for each of the new 11 subjects. It can be seen that for 14 out of 15 subjects the classification accuracy of all the channels and the occipitoparietal channels can be improved by using subsets of channels that belong to the Pareto set obtained for four of the subjects. For some subjects the accuracy improvement can be over 10 %. What we want to emphasize at this point

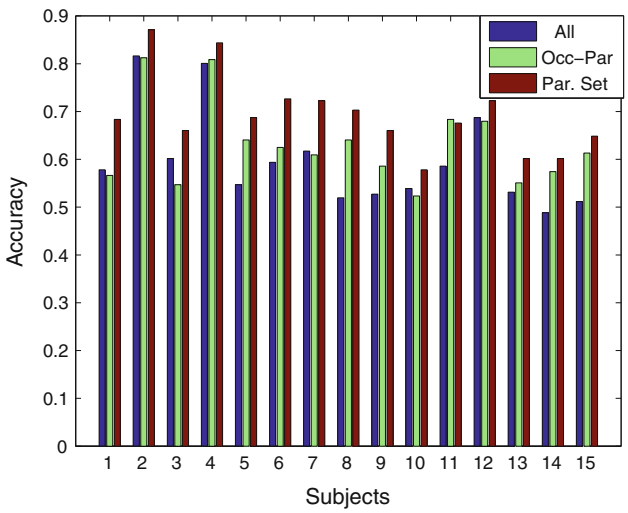

Fig. 10 Classification accuracy achieved in the complete set of 15 subjects and with the raw+correlation type of information. Different subsets of channels have been used: all the channels, the occipitoparietal channels, and the best solution of the Pareto-set computed using only 4 subjects

is that the Pareto set of solutions found for a reduced set of subjects can be used as a reservoir of potential solutions for other subjects. We can also think of situations where brain signals from many subjects are available and an initial clustering step is applied to select a subset of exemplar subjects for which the channel multiobjective optimization step will be conducted subsequently.

5.6 Analyzing the parameters of the logistic models

Through channel selection the classifier is able to reduce the data used for classification, focusing on the most task-related informative brain areas. However, the information coming from all the channels is not necessarily equally relevant for the classifier in informative terms. The relevance of a channel will depend on the contribution of the features associated to each channel. Notice that selecting a channel means that all the features coming from that channel will be used for classification (e.g., for raw information there are 60 features for each channel). One way to assess the relevance of a channel is to inspect the logistic model parameters associated with the feature variables coming from it. Regularized classifiers tend to set the parameters corresponding to those variables that are not relevant for classification to zero.

We compute the frequencies with which the parameters corresponding to each feature have been set to zero in all the solutions (set of channels) that belong to the Pareto set of solutions. The most relevant features from each channel are expected to be those most frequently selected by the classifier across solutions. Additionally, we do not expect the channels that have the parameters for all their variables set to zero to be important for classification. Due to the complexity of the optimization problem, it may occur that channels whose features do not contribute to the classification are selected by the optimization algorithm. Therefore, determining the feature relevance can be seen as a refinement of our channel selection method. It will improve the quality of the extracted biological information, discarding channels whose corresponding features are seldom identified by the classifier as relevant.

We analyze the set of 152 Pareto solutions found for the raw+correlation type of information. For this type of information a channel i may be relevant because either the variables representing raw information from channel i or the variables representing correlation information that involves channel i are frequently nonzero in the classifiers. To analyze the different sources of relevance, we separate the analysis of these two cases. Figure 11 shows the channels that, on the basis of the analysis of their corresponding nonzero coefficients in the regularized models, are identified as relevant because their contribution comes from the raw information. These channels are represented by black dots. The colors indicate the average absolute value of coefficients.

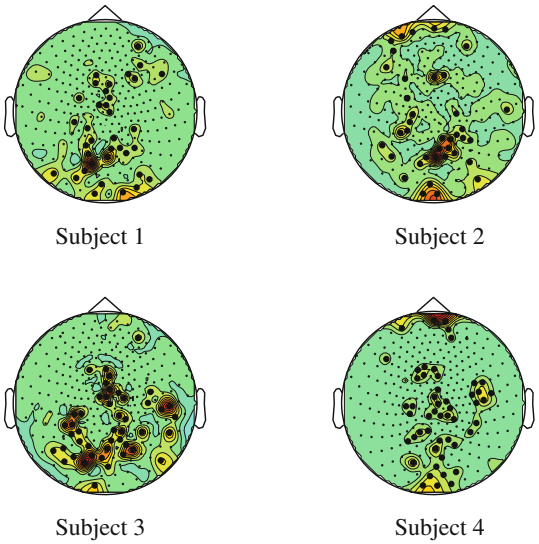

Fig. 11 Channels that were identified as relevant due to their contribution coming from the raw information, as a result of the analysis of their corresponding non-zero coefficients in the regularized models. The raw+correlation type of information was used

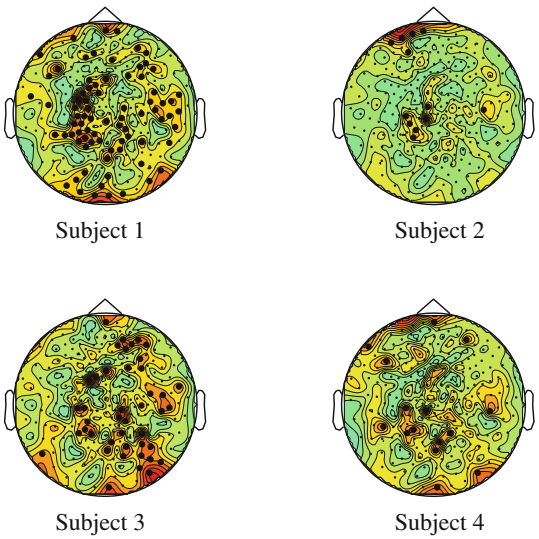

Fig. 12 Channels that were identified as relevant due to their contribution coming from the correlation information, as a result of the analysis of their corresponding non-zero coefficients in the regularized models. The raw+correlation type of information was used

Similarly, Fig. 12 shows the channels identified as relevant because their contribution comes from the correlation information components. In this case, the figure represents the channels involved in each pairwise correlation detected as relevant. For the two cases considered, we set a threshold for the minimal number of times that the coefficients corresponding to the features of each channel should be different from zero. This threshold is 80 for the raw type of information and

50 for the correlation type of information. Only those channels whose coefficients satisfy these constraints are shown in Figs. 11 and 12.

Figure 11 shows that the number of relevant channels is less than the number of selected channels of the single solution shown in Fig. 8. Selected channels are also more related to the information available a priori about brain areas involved in the studied brain processes. By contrasting Figs. 11 and 12, we recognize two different situations concerning the variable contribution of raw data and the channel correlations to classification accuracy.

The first situation, illustrated by subject 1, is when the number of channels whose contribution is due to the correlations is higher than the number of channels whose contribution is determined by the raw data. The second situation, illustrated by subject 2, is the opposite. For this subject, only a few of the channels have coefficients associated with the correlation features that are above the fixed threshold. What we want to emphasize here is that the information captured by logistic regression models may be useful for classifying individuals according to the different dynamics involved in their mental activity. This classification could be useful to explain intersubject variability and eventually to tailor BCI to the particular characteristics of the subjects.

It is also interesting to look at the periods of the recorded time series that are more informative for the classification task. Since the raw data corresponding to each channel are codified using 60 variables, we can look at the coefficients learned by the models for each of these variables at each channel. Fixing a threshold based on the number of nonzero coefficients produced by each variable, we can give estimates about the relevance of each time period. We have computed the number of nonzero coefficients for each of the 60 variables in all the solutions of the Pareto set for the raw+correlation type of information and for each subject. Figure 13 shows the total number of nonzero coefficients

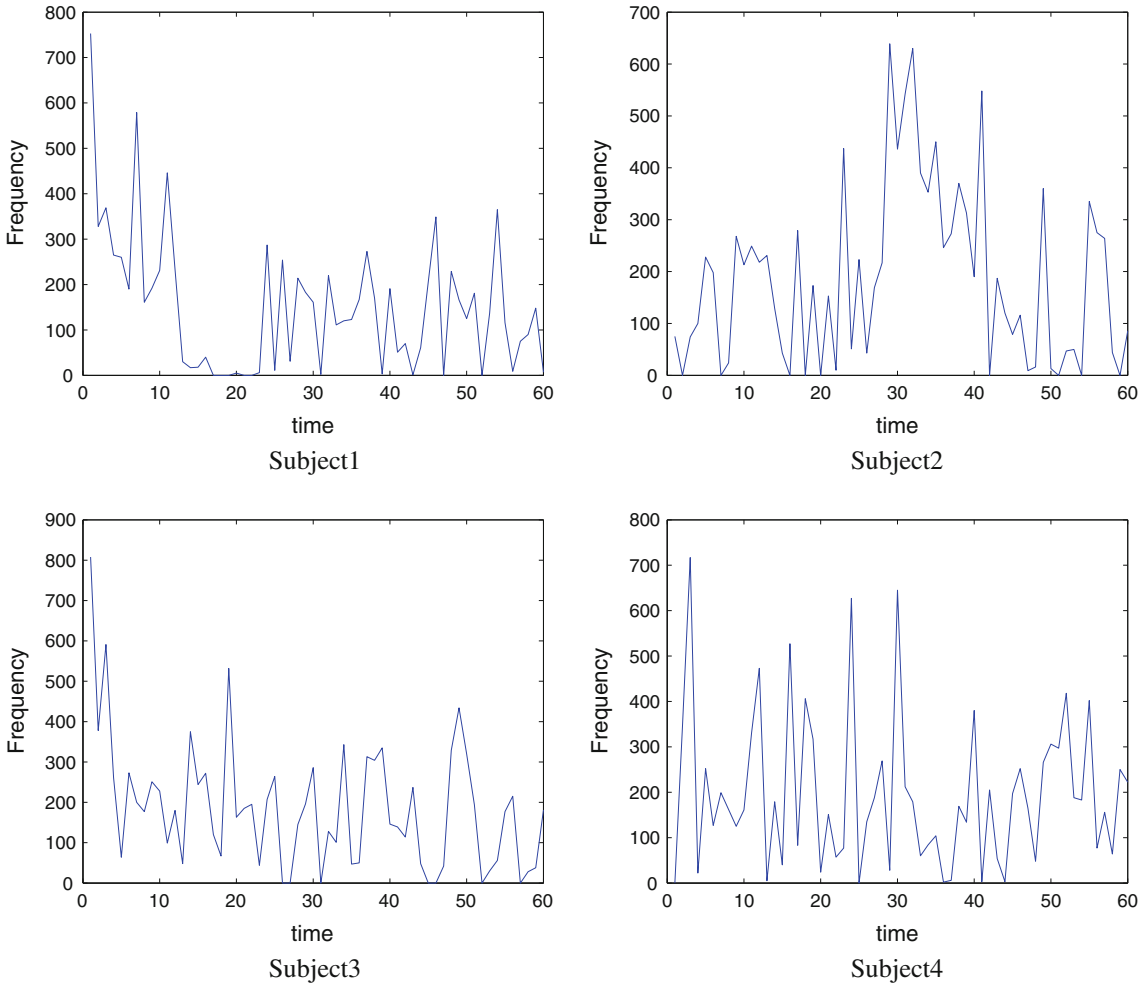

Fig. 13 Number of non-zero coefficients for each of the 60 variables describing the raw information component of the raw+correlation information and summed from all channels represented by the 152 solutions of the Pareto set

computed from all the Pareto solutions and adding all the channels.

Figure 13 shows a very strong contribution of nonzero coefficients during the initial periods of the time series for subjects 1, 3, and 4. For these subjects the highest contributions come from the first variables. A different trend is observed for subject 2, where the highest contributions are in the middle of the recorded time series. A channel-specific, differential analysis of the informative value of the time variables can be expected to provide a better understanding of this question. This type of analysis, which could serve to reveal physiologically relevant information in the data, is left for future work.

6 Related work

Several approaches that apply regularization methods to extract information from MEG and EEG have been proposed Haufe et al. (2010); Valdes-Sosa et al. (2005); van-Gerven and Jensen (2009). The approach most related to our work is presented in van-Gerven et al. (2009), where sparse logistic regression using lasso regularization is used to solve the same classification problem but using a different type of predicting variable. In Obermaier et al. (2001), GAs are combined with hidden Markov models (HMMs) for classification in an offline EEG-based BCI. The authors found that the use of asymmetrical classifiers derived from the GA-based proposal performed significantly better than the HMM classifier.

A number of methods based on the analysis of time series extracted from MEG have been proposed for assessing functional connectivity between brain regions (Darvas and Leahy 2007). These include, for example, the use of covariance, mutual information, coherence. de Lange et al. (2008) use cross-frequency amplitude coupling to identify interactions between different brain areas during imagined actions. It is important to emphasize that just because two or more regions share mutual information during a given task does not imply a causal relationship between these regions. Furthermore, correlations between the information collected by channels may be due to artifacts in the registering procedure and not to neuron activity.

Importantly, the use of network topological measures to analyze graphs constructed from MEG data is not new. In Bucolo et al. (2008) and DiGrazia et al. (2009), the synchronization likelihood, a statistical measure of dependence between channels, is used to construct an interaction graph to investigate the occurrence of small-world phenomena in MEG data. Bucolo et al. and DiGrazia et al. extracted topological measures (clustering coefficient, path length, mean degree) to characterize the differences in the three different phases of the evaluated experimental protocol.

7 Conclusions and future work

In this paper we have proposed a unified approach that combines a number of methods to improve classification accuracy when covert spatial attention is used for BCI. We assert that by combining raw information with features extracted from the time series correlations, it is possible to achieve a higher accuracy than by using only one type of information. To further improve the results accuracy, we have shown that multiobjective optimization using EAs is a valid alternative for selecting accurate subsets of channels. Using the output Pareto set of solutions, we conducted a global analysis of the classification problem. Instead of focusing on a single solution, we showed that by inspecting the Pareto sets, it is possible to unveil knowledge about the channels that are more frequently involved in accurate classification.

From our results we have confirmed that regularized logistic regression is a very suitable classifier. It increases the classifier generalization capabilities and incorporates numerous features into the classification task. This is a particularly important characteristic for MEG analysis since usually information coming from hundreds of channels is available. The use of surrogate accuracy values in the form of two-fold cross-validation accuracies during the EA evolution has proved to reduce the cost of the evolutionary search and pointed the search in the right direction. Also, the idea of extending the information passed to the classifier by adding features that were not originally included in the search for the optimal classifier has shown that it is possible to improve accuracy with no added cost associated with the search.

Although current BCIs use electrophysiological signals representing brain events that are reasonably well defined anatomically and physiologically (Wolpaw et al. 2002), the exploration of signal features that exhibit more complex relationships to the underlying difficult-to-explain brain events remains a promising path for BCI development. Machine learning techniques used for classification are appropriate for this task. They should be able to provide compact but still legible characterizations (models) of the signal features. Furthermore, it should be possible to identify distinctive patterns in the brain dynamics of different subjects from the analysis of these models.

Acknowledgments We would like to thank the Donders Centre for Cognitive Neuroimaging for providing the data used in our experiments and to Marcel van Gerven for useful comments on this data. This work has been partially supported by the Saiotek and Research Groups 2007–2012 (IT-242-07) programs (Basque Government), Cajal Blue Brain, TIN-2008-06815-C02-02, TIN2010-20900-C04-04, TIN2010-14931 and Consolider Ingenio 2010-CSD 2007-00018 projects (Spanish Ministry of Science and Innovation) and COMBIOMED network in computational biomedicine (Carlos III Health Institute).

References

Armañanzas R, Saeys Y, Inza I, García-Torres M, Bielza C, van de Peer Y, Larrañaga P (2011) Peakbin selection in mass spectrometry data using a consensus approach with estimation of distribution algorithms. IEEE/ACM Trans Comput Biol Bioinf 8(3):760–774

Asano F, Kimura M, Sekiguchi T, Kamitani Y (2009) Classification of movement-related single-trial MEG data using adaptive spatial filter. In: Acoustics, speech and signal processing, 2009. ICASSP 2009. IEEE international conference on, pp 357–360. IEEE, Rotterdam

Besserve M, Jerbi K, Laurent F, Baillet S, Martinerie J, Garnero L (2007) Classification methods for ongoing EEG and MEG signals. Biol Res 40(4):415–437

Bianchi L, Sami S, Hillebrand A, Fawcett I, Quitadamo L, Seri S (2010) Which physiological components are more suitable for visual ERP based brain–computer interface? A preliminary MEG/EEG study. Brain Topogr 23(2):180–185

Bucolo M, Di Grazia F, Frasca M (2008) From synchronization to network theory: a strategy for MEG data analysis. In: Proceedings of 16th mediterranean conference on control and automation, pp 854–859, Ajaccio, France. IEEE Press, Piscataway, NJ

Carmena J, Lebedev M, Crist R, Doherty J, Santucci D, Dimitrov D, Patil P, Henriquez C, Nicolelis M (2003) Learning to control a brain-machine interface for reaching and grasping by primates. PLoS Biol 1(2):193–208

Coello C, Lamont G, Van Veldhuizen D (2007) Evolutionary algorithms for solving multi-objective problems. Springer, New York

Darvas F, Leahy RM (2007) Handbook of brain connectivity, chapter functional imaging of brain activity and connectivity with MEG, pp 201–220. Kluwer Academic Publishers, Boston

de Lange FP, Jensen O, Bauer M, Toni I (2008) Interactions between posterior gamma and frontal alpha/beta oscillations during imagined actions. Front Hum Neurosci 2(7):1–12

Di Grazia F, Sapuppo F, Shannahoff-Khalsa D, Bucolo M (2009) Network parameters for studying functional connectivity in brain MEG data. Int J Bioelectromagn 11(4):161–169

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22

Goldberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Reading, MA

Guimera R, Amaral LAN (2005) Functional cartography of complex metabolic networks. Nature 433:895–900

Haufe S, Tomioka R, Nolte G, Muller K, Kawanabe M (2010) Modeling sparse connectivity between underlying brain sources for eeg/meg. IEEE Trans Biomed Eng 57(8):1954–1963

Hoffmann U, Vesin J, Ebrahimi T, Diserens K (2008) An efficient p300-based brain-computer interface for disabled subjects. J Neurosci Methods 167(1):115–125

Holland JH (1975) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. University of Michigan Press, Ann Arbor, MI

Inza I, Larrañaga P, Etxeberria R, Sierra B (2000) Feature subset selection by Bayesian network-based optimization. Artif Intell 123 (1–2):157–184

Iturrate I, Antelis J, Minguez J, Kübler A (2009) A non-invasive brain-actuated wheelchair based on a p300 neurophysiological protocol and automated navigation. IEEE Trans Robot 25(3):614–627

Kelly S, Lalor E, Finucane C, McDarby G, Reilly R (2005) Visual spatial attention control in an independent brain-computer interface. IEEE Trans Biomed Eng 52(9):1588–1596

Larrañaga P, Lozano JA (eds) (2002) Estimation of distribution algorithms. A new tool for evolutionary computation. Kluwer Academic Publishers, Boston

Lebedev M, Nicolelis M (2006) Brain-machine interfaces: past, present and future. Trends Neurosci 29(9):536–546

Leicht EA, Newman MEJ (2008) Community structure in directed networks. Phys Rev Lett 100:118703

Lotte F, Congedo M, Lecuyer A, Lamarche F, Arnaldi B (2007) A review of classification algorithms for EEG-based brain–computer interfaces. J Neural Eng 4:R1–R13

McLachlan G (1992) Discriminant analysis and statistical pattern recognition. Wiley, New York

Mendiburu A, Miguel-Alonso J, Lozano JA, Ostra M, Ubide C (2006) Parallel EDAs to create multivariate calibration models for quantitative chemical applications. J Parallel Distrib Comput 66(8):1002–1013

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298:824–827

Mühlenbein H, Paaß G (1996) From recombination of genes to the estimation of distributions I. Binary parameters. In: Voigt H-M, Ebeling W, Rechenberg I, Schwefel H-P (eds) Proceedings of the 4th international conference on parallel problem solving from nature-PPSN IV, vol 1141 of lectures notes in computer science, pp 178–187. Springer, Berlin

Nicolelis M (2003) Brain–machine interfaces to restore motor function and probe neural circuits. Nat Rev Neurosci 4(5):417–422

Obermaier B, Munteanu C, Rosa A, Pfurtscheller G (2001) Asymmetric hemisphere modeling in an offline brain–computer interface. IEEE Trans Syst, Man, Cybern C 31(4):537–540

Pelikan M, Goldberg DE, Lobo F (2002) A survey of optimization by building and using probabilistic models. Comput Opt Appl 21(1):5–20

Rieger J, Reichert C, Gegenfurtner K, Noesselt T, Braun C, Heinze H, Kruse R, Hinrichs H (2008) Predicting the recognition of natural scenes from single trial MEG recordings of brain activity. Neuroimage 42(3):1056–1068

Rossini L, Izzo D, Summerer L (2009) Brain-machine interfaces for space applications. In: Proceedings of the annual international conference of the IEEE engineering in medicine and biology society, vol 1, pp 520–523, Minnesota

Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517

Santana R, Bielza C, Larrañaga P (2010a) Synergies between network-based representations and probabilistic graphical modeling in the solution of problems from neuroscience. In: García-Pedrajas N et al. (eds) Proceedings of the twenty third international conference on industrial, engineering and other applications of applied intelligent systems, vol 6098 of lecture notes in artificial intelligence, pp 149–158, Springer, Córdoba

Santana R, Bielza C, Larrañaga P (2010b) Using probabilistic dependencies improves the search of conductance-based compartmental neuron models. In C. Pizzuti, M. D. Ritchie, and M. Giacobini, editors, Proceedings of the 8th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, volume 6023 of Lecture Notes in Artificial Intelligence, pages 170–181. Springer

Santana R, Bielza C, Larrañaga P, Lozano JA, Echegoyen C, Mendiburu A, Armañanzas R, Shakya S (2010c) Mateda-2.0: A MATLAB package for the implementation and analysis of estimation of distribution algorithms. J Stat Softw 35(7):1–30

Santana R, Ochoa A, Soto MR (2001) The mixture of trees factorized distribution algorithm. In: Proceedings of the genetic and evolutionary computation conference GECCO-2001, pp 543–550. Morgan Kaufmann Publishers, San Francisco, CA

Sporns O (2002) Neuroscience databases. A practical guide, chapter graph theory methods for the analysis of neural connectivity patterns, pp 171–186. Kluwer, Boston, MA

Tan L, Jansari A, Keng S, Goh S (2009) Human-computer interaction. Novel interaction methods and techniques, chapter effect of mental training on BCI performance, pp 632–635. Springer, Berlin

The MathWorks Inc. (2007) MATLAB—the language of technical computing, version 7.5. The MathWorks Inc., Natick, MA

Valdés-Sosa PA, Sánchez-Bornot JM, Lage-Castellanos A, Vega-Hernández M, Bosch-Bayard J, Melie-García L, Canales-Rodríguez E (2005) Estimating brain functional connectivity with sparse multivariate autoregression. Philos Trans 360(1457): 969–981

van-Gerven M, Bahramisharif A, Heskes T, Jensen O (2009) Selecting features for BCI control based on a covert spatial attention paradigm. Neural Networks 22:1271–1277

van-Gerven M, Jensen O (2009) Attention modulations of posterior alpha as a control signal for two-dimensional brain-computer interfaces. J Neurosci Methods 179:78–84

Vapnik V (2000) The nature of statistical learning theory. Springer, New York

Waldert S, Braun C, Preissl H, Birbaumer N, Aertsen A, Mehring C (2007) Decoding performance for hand movements: EEG vs. MEG. In: Engineering in medicine and biology society, 2007. EMBS 2007. 29th Annual international conference of the IEEE, pp 5346–5348. IEEE, Washington, DC

Wang W, Sudre G, Xu Y, Kass R, Collinger J, Degenhart A, Bagic A, Weber D (2010) Decoding and cortical source localization for intended movement direction with MEG. J Neurophysiol 104(5):2451–2461

Wolpaw J, Birbaumer N, McFarland D, Pfurtscheller G, Vaughan T (2002) Brain-computer interfaces for communication and control. Clin Neurophysiol 113(6):767–791

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc: Ser B 67(2):301–320