Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 350270, 11 pages
http://dx.doi.org/10.1155/2014/350270

Research Article

EEG Channel Selection Using Particle Swarm Optimization for the Classification of Auditory Event-Related Potentials

Alejandro Gonzalez, Isao Nambu, Haruhide Hokari, and Yasuhiro Wada

Department of Electrical Engineering, Nagaoka University of Technology, 1603-1 Kamitomioka, Nagaoka, Niigata 940-2188, Japan

Correspondence should be addressed to Alejandro Gonzalez; alejandrojgt@stn.nagaokaut.ac.jp

Received 6 January 2014; Accepted 26 February 2014; Published 25 March 2014

Academic Editors: V. Bhatnagar and Y. Zhang

Brain-machine interfaces (BMI) rely on the accurate classification of event-related potentials (ERPs) and their performance greatly depends on the appropriate selection of classifier parameters and features from dense-array electroencephalography (EEG) signals. Moreover, in order to achieve a portable and more compact BMI for practical applications, it is also desirable to use a system capable of accurate classification using information from as few EEG channels as possible. In the present work, we propose a method for classifying P300 ERPs using a combination of Fisher Discriminant Analysis (FDA) and a multiobjective hybrid real-binary Particle Swarm Optimization (MHPSO) algorithm. Specifically, the algorithm searches for the set of EEG channels and classifier parameters that simultaneously maximize the classification accuracy and minimize the number of used channels. The performance of the method is assessed through offline analyses on datasets of auditory ERPs from sound discrimination experiments. The proposed method achieved a higher classification accuracy than that achieved by traditional methods while also using fewer channels. It was also found that the number of channels used for classification can be significantly reduced without greatly compromising the classification accuracy.

1. Introduction

A brain-machine interface (BMI) is a system that allows a person to control or communicate with a computer or actuator using only brain signals. Since no muscle movements are needed, such systems are particularly useful to assist patients with motor disabilities such as amyotrophic lateral sclerosis or spinal cord injury.

One type of BMI system makes use of the P300 event-related potential (ERP), a neuroelectrical wave pattern that can be measured with electroencephalography (EEG). This is a pattern that carries information about the state of attention of the user and that can be robustly elicited by various types of stimuli through the oddball paradigm. In the oddball paradigm, the user pays attention to a series of incoming stimuli and must focus on the occurrences of a rare, task-relevant target stimulus hidden amongst frequent, irrelevant nontarget stimuli. Only the target stimuli elicit the P300 response when perceived by the user. The nontarget stimuli, on the other hand, are more frequent and are ignored by the user and thus do not elicit the P300 response. Therefore, by presenting to the user a stream of stimuli while measuring the user's brain the activity and then detecting the presence or absence of the P300 ERP, one can determine the intention of the user.

An example of this kind of system is the one proposed and investigated in [1–3]. This system allows a handicapped patient to communicate to a computer the direction of attention by detecting the P300 ERPs elicited with auditory stimuli from virtual-sound sources. By using the virtual sounds as the stimuli in the oddball paradigm, it is possible to estimate the intended direction of the user and thus control a transportation device like an electric wheelchair. The use of virtual sound stimuli allows the construction of more portable BMI systems and, in contrast to visual stimuli, allows the user to dedicate his or her vision to other tasks.

The detection of the P300 component is performed by means of a binary classifier fed with the signal features provided by the EEG signals, and its performance greatly depends on the chosen features. Ideally only the most discriminative features should be used to feed the classifier but, in this case, it is usually not immediately clear which is the channel set that provides the most relevant information. Furthermore, the optimal set might not be equal for every subject. One could attempt to perform an exhaustive test of all possible combinations, but, in the case of dense array measurements, the extremely high number of combinations that exist (2^{64}, for a 64-channel array) render this approach intractable. Additionally, in practical applications, a BMI system for transportation purposes like the one described in [3] requires compact equipment and it is thus desirable to use a ERP detection system that employs as few channels as possible.

Stepwise Linear Discriminant Analysis (SWLDA) is one of the most popular classifiers used for the detection of ERPs, being used in numerous reports (see, e.g., [4–7]). This classifier is based on the feature selection performed by the Stepwise Regression algorithm in which the features that contribute the most are selected in a stepwise manner; that is, every feature is sequentially added or removed while measuring the predicting power at each step. However, a drawback of this algorithm is that the selection of the features depends on the order on which they are evaluated, especially when there are high correlations between the features [8] which is precisely the case in EEG measurements. Another algorithm worth mentioning is the one proposed by [9] that achieved the best performance in the BCI Competition III [10]. This algorithm performs a stepwise selection of channels following a classification accuracy maximization criterion, but, like SWLDA, the outcome of the procedure also depends on the order of the evaluation of channels. There are also algorithms that employ methods based on Principal Component Analysis (PCA) to extract spatial features [6, 11, 12], but these typically disregard channel selection and employ all the channels available to identify the most discriminative spatial information.

To address the abovementioned points, we propose an alternative approach in which the channel selection procedure is performed in an automated manner aimed at maximizing the classification accuracy of the system. Specifically, we propose a method for classifying P300 ERPs in which the features and the parameters of the classifier are tuned using a random optimization algorithm and evaluate it using experimental data. The proposed method performs classification using Fisher Discriminant Analysis (FDA) and uses a multiobjective hybrid real-binary Particle Swarm Optimization (MHPSO) algorithm to search for the classifier parameters and EEG channel set that simultaneously maximize the classification accuracy and minimize the number of channels used for classification. The PSO algorithm is multiobjective in the sense that it seeks to optimize a fitness function product of an aggregation of two performance metrics (i.e., the classification accuracy and the number of EEG channels) and hybrid real-binary in the sense that the search space is composed of both real and binary dimensions, necessary to tune the FDA regularization parameter (a real variable) and the addition or removal of EEG channels (a binary variable per channel).

PSO [13] is a relatively new stochastic algorithm for function optimization that has become increasingly popular and has also been applied in various fields [14]. This algorithm, inspired in the motion of fish schools that search for food, moves a swarm of particles around a parameter space searching for a solution that maximizes a certain fitness function. Like other stochastic search algorithms, this is a method that does not rely on the gradient of the function to optimize and instead looks for the best solution in a quasirandom way, being particularly useful in problems where an analytical expression for the optimization function is not available. It is important to remark that the PSO algorithm is not guaranteed to converge towards the global maximum. However, this algorithm has been applied with success in a wide variety of applications [14] and in practice tends to find a suitable solution if not the optimal one.

This text is organized as follows. First, the main components of the proposed algorithms, namely, the FDA and PSO algorithms, and their integration to classify ERP signals are described in Section 2. Then, the data and simulations used to evaluate the algorithm are described in Section 3. Lastly, in Section 4, the simulations results are shown and discussed and conclusions are given in Section 5.

2. Methods

Note on Mathematical Notation. Throughout this work, scalars are denoted by lowercase italic letters (e.g., x), vectors are written in lowercase, bold letters (e.g., $\mathbf{x}$), and matrices are denoted by uppercase bold letters (e.g., $\mathbf{X}$). All vectors are column vectors unless stated otherwise.

2.1. Particle Swarm Optimization. The objective of the PSO algorithm is to find the best parameters that maximize a given fitness function, and it does so by iteratively moving a swarm of particles around the parameter space according to especial equations. Each particle of the swarm is a particular position in the parameter space and represents a possible solution to the problem. Mathematically, a particle in PSO is a vector in an N-dimensional parameter space, and its position $\mathbf{x}$ and velocity $\mathbf{v}$ change according to the following equations:

$$x_{i,j}^{t} = x_{i,j}^{t-1} + v_{i,j}^{t}, \tag{1}$$

$$v_{i,j}^{t} = w v_{i,j}^{t-1} + c_1 \eta_1 \left(p_{i,j} - x_{i,j}^{t-1} \right) + c_2 \eta_2 \left(g_j - x_{i,j}^{t-1} \right), \tag{2}$$

where $x_{i,j}^{t}$ and $v_{i,j}^{t}$ are the jth components of the ith particle's position and velocity, respectively, at iteration t. $p_{i,j}$ is the jth component of $\mathbf{p}$, the best position that the ith particle has found so far, and g_j is the jth component of $\mathbf{g}$, the best position found by the swarm. The effect of $\mathbf{p}$ and $\mathbf{g}$ on the particle's motion is controlled by two constant parameters c_1 and c_2, and two independent random variables η_1 and η_2 uniformly distributed in $[0, 1]$. The particle's motion is also influenced by the velocity at the previous iteration and this effect is controlled by the inertia parameter w.

c_1 and c_2 are constants set by the experimenter that determine the balance between the exploitation of a potential solution (movement towards **g**) and the exploration for new solutions (movement towards **p**). At each iteration, **p** and **g** are updated if positions with better fitness were found. This is the main feature of the PSO algorithm: each particle uses the information of its own history and the swarm's history, together with random perturbations, to search for the global maximum. The PSO algorithm, in an iterative manner, updates the velocity and positions of each particle, moving them around the parameter space, until the best global solution **g** reaches the desired fitness.

In practical applications, the search space is typically constrained to $[X_j^{\min}, X_j^{\max}]$ along each dimension j in order to limit the search space to feasible values. In this work, we enforced an invisible wall condition [15] in which the fitness of the particles that fly out of this region is neither calculated nor updated. Instead, the particles that stray out are expected (but not guaranteed) to eventually fly back into the admissible region. The invisible wall condition has the advantage of being simple to implement and avoiding the unnecessary computations required by the evaluation of unfeasible solutions. Lastly, to avoid the particles flying out of the feasible space too often, the velocity components are limited to a maximum value $V_j^{\max}$ such that

$$\left|v_{i,j}^t\right| < V_j^{\max}. \tag{3}$$

Hybrid Binary-Real PSO. In order to perform channel selection, in this work, we adopted a hybrid binary-real PSO algorithm that, in addition to real-valued variables, allows the optimization of variables that can take only two discrete values. In this case, the additional binary components are updated as shown below [15–17]:

$$x_{i,j}^t = \begin{cases} 1 & \text{if } r < S\left(v_{i,j}^t\right) \\ 0 & \text{if } r \geq S\left(v_{i,j}^t\right) \end{cases}, \quad S(x) = \frac{1}{1+e^{-x}}, \tag{4}$$

where r is a random number with a uniform distribution in $[0,1]$ and the velocity components $v_{i,j}^t$ are updated in a manner similar to the real case (see (2)). This equation means that, at each iteration, the binary component $x_{i,j}^t$ will be 1 (i.e., the associated channel will be selected) with a probability of $S(v_{i,j}^t)$ (and 0 with a probability of $1 - S(v_{i,j}^t)$).

2.2. Fisher Discriminant Analysis.

In this work, a linear binary classifier based on Fisher Discriminant Analysis (FDA) was applied to discriminate between target and nontarget signals. FDA is a machine learning technique proposed by [18] used for data classification. Strictly speaking, FDA is a dimensionality reduction used as a preprocessing step prior to classification and its goal is to find a linear combination of the features that maximizes the separation between the classes' distributions in the reduced space. Classification is then performed in this 1-dimensional space by applying some threshold criteria or by using any classifier trained on the transformed samples. Given a training dataset of n-dimensional samples $\{(\mathbf{z}_i, y_i),\ i = 1, \ldots, \ell\}$ (i.e., each sample has n features) were each of the samples belonging to either one of two classes K_1 (positive, e.g., target examples) or K_{-1} (negative, e.g., nontarget examples) as indicated by the categorical variable $y_i = \{+1, -1\}$, the transformation vector **w** is obtained by maximizing the following function:

$$J(\mathbf{w}) = \frac{\left\langle \mathbf{w}, \mathbf{m}_1 - \mathbf{m}_{-1} \right\rangle^2}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}. \tag{5}$$

($\langle \cdot, \cdot \rangle$ denotes the inner product of vectors,) $\mathbf{m}_1$ and $\mathbf{m}_{-1}$ are the means of the feature vectors of the positive and negative classes, respectively, and $\mathbf{S}_W$ is the within-class scatter matrix given by

$$\mathbf{S}_W = \sum_{j \in \{1,-1\}} \sum_{\mathbf{z}_i \in K_j} \left(\mathbf{z}_i - \mathbf{m}_j\right)\left(\mathbf{z}_i - \mathbf{m}_j\right)^T, \tag{6}$$

with

$$\mathbf{m}_1 = \frac{1}{\ell_1} \sum_{\mathbf{z}_i \in K_1} \mathbf{z}_i, \qquad \mathbf{m}_{-1} = \frac{1}{\ell_{-1}} \sum_{\mathbf{z}_i \in K_{-1}} \mathbf{z}_i. \tag{7}$$

The maximization of $J(\mathbf{w})$ therefore yields

$$\mathbf{w} = \mathbf{S}_W^{-1}\left(\mathbf{m}_1 - \mathbf{m}_{-1}\right), \tag{8}$$

and the classification of a new, unknown sample, **z**, is performed upon the score

$$s(\mathbf{z}) = \langle \mathbf{w}, \mathbf{z} \rangle, \tag{9}$$

which is the distance of the sample to the hyperplane parameterized by the normal vector **w** and represents a measure of the certainty about the class prediction. Regarding the class prediction, in this work, the class of the sample was determined by

$$\hat{y} = \arg\min_{j \in \{1,-1\}} \sqrt{\frac{\left(s(\mathbf{z}) - \mu_j\right)^2}{\sigma_j^2}}, \tag{10}$$

where μ_j and σ_j are the mean and variance of the scores of the training samples belonging to class K_j $(j \in \{1,-1\})$ in the transformed space. Equation (10) means that the predicted class corresponds to the class that yields the shortest Mahalanobis distance to the sample.

The key element in the computation of **w** is the scatter matrix $\mathbf{S}_W$. $\mathbf{S}_W$ is an unbiased estimator of the true, unknown scatter matrix, and it may become imprecise when the number of features is high in comparison to the number of training samples. This is because the number of unknown parameters (the elements of the matrix) is quadratic in the number of features. An imprecise estimation of the within-class scatter matrix results in a degradation of classification performance [5]. To mitigate this effect, regularization is typically applied to the scatter matrix estimation and this is achieved by maximizing a modified target function $J(\mathbf{w})$ [19]:

$$J(\mathbf{w}) = \frac{\left\langle \mathbf{w}, \mathbf{m}_1 - \mathbf{m}_{-1} \right\rangle^2}{\mathbf{w}^T \mathbf{S}_W \mathbf{w} + \lambda \|\mathbf{w}\|^2}, \tag{11}$$

which results in

$$\mathbf{w} = \left(\mathbf{S}_W + \lambda \mathbf{I}\right)^{-1} \left(\mathbf{m}_1 - \mathbf{m}_{-1}\right), \tag{12}$$

where λ is the regularization parameter. It can be seen that for $\lambda = 0$ the canonical, unregularized FDA is obtained. Thus, it is necessary to appropriately choose the regularization parameter in order to achieve the best performance.

2.3. The Proposed Algorithm

2.3.1. Particle Encoding. As was mentioned before, the algorithm uses FDA for classification and a multiobjective hybrid PSO to tune, for each particular user, the channel set and classifier parameters that maximize the classification accuracy using as few channels as possible.

The proposed method employs a hybrid PSO algorithm to search in a space that contains both real and discrete (binary) dimensions in a similar fashion to [15, 17]. These dimensions correspond to the variables that are to be tuned. Thus, each particle is a position in the search space that represents a particular combination of FDA parameters and EEG channels and is a candidate solution for the problem. The channel set and classifier parameters are encoded in a particle $\mathbf{x}$ as follows:

$$\mathbf{x} = \left[a \ \ b_1 \ \ \cdots \ \ b_{64}\right], \tag{13}$$

where $a \in [-1, 1]$ and $b_j = \{0, 1\}\ \forall j$ denote the real and binary components, respectively. Each binary component (64 in total) b_j encodes whether the temporal features of the corresponding jth channel are used for classification ($b_j = 1$) or not ($b_j = 0$). Therefore, this encoding results in a search space of 65 dimensions. The real component a encodes the FDA regularization parameter λ, which is decoded as

$$\lambda = 10^{5a}. \tag{14}$$

This decoding was chosen because the feasible values for the regularization parameter usually range over several orders of magnitude, and by adopting an exponential representation, the particle can search with small steps for possible solutions over a broader interval.

2.3.2. Fitness Function. As mentioned above, each particle is a candidate solution and represents a possible combination of set of channels and FDA parameters. In this work, we search for the solution that yields the best performance measured in terms of the fitness function F:

$$\begin{aligned} F(\mathbf{x}) &= w_1 f_1(\mathbf{x}) + w_2 f_2(\mathbf{x}) \\ &= w_1 \sqrt{\frac{TP}{P_s} \times \frac{TN}{N_s}} + w_2 \left(\frac{N_{\mathrm{Ch}} - n + 1}{N_{\mathrm{Ch}}}\right), \end{aligned} \tag{15}$$

where TP stands for true positives, TN stands for true negatives, P_s and N_s are the total number of positive (target) and negative (nontarget) samples, respectively, N_{Ch} is the maximum number of channels that can be selected (64 in this work), and n is the number of channels in the set specified by the particle. In other words, F is a weighted aggregation of the optimization objectives f_1 and f_2: the maximization of the geometric mean between the true positive rate (target accuracy) and the true negative rate (nontarget accuracy) and the minimization of the number of channels. The trade-off between accuracy and number of channels is controlled by the fitness weights w_1 and w_2. Since the ratio between the weights is what actually determines the trade-off, in this work, we chose weights such that $w_1 + w_2 = 1$. The geometric mean was chosen over the arithmetic mean because it yields an aggressive evaluation in which solutions with low and/or highly unbalanced accuracies are heavily penalized. It is important to notice that the target and nontarget accuracies are not expressed as percentages and thus f_1 varies within $[0, 1]$, being 1 the perfect accuracy. f_2, on the other hand, yields higher values for fewer channels and becomes 1 when only one channel is selected. It is important to remark that, although the algorithm attempts to find a solution that yields a perfect classification using only one channel (fitness value equal to 1), such a solution might not exist because the information provided by one channel might not be enough to accurately discriminate the ERP signals.

In the offline analyses, for each particle, a classifier is built using the features and parameters encoded in the particle. Then, the true positive and true negative rates achieved by the classifier are estimated using 10-fold stratified cross-validation on the training data (all folds share the same target to nontarget sample ratio), and the particle fitness is finally calculated using (15). The proposed algorithm's flowchart is shown in Figure 1. A final classifier can then be trained using the best setup found by the algorithm and be used to classify signals in a real-time setting.

3. Simulations

3.1. Data Set Used in the Study. The data used in this work corresponds to the data gathered by [1, 2] of sound discrimination experiments that consisted of random presentations of virtual auditory stimuli from 6 directions. During these experiments, the subject had to focus his attention on one of these directions (the target direction) and count every time the stimulus source corresponded to the target direction. The subject had to ignore the stimuli from other directions (the nontarget directions). Each session consisted of 150 trials and each trial consisted of a 300 ms stimulus interval and an 800 ms silence interval. The stimulus was 300 ms of pure white noise. One of the 6 directions was fixed as the target direction throughout a single session and the corresponding stimuli were presented with a probability of 20%. Each direction was measured twice, thus yielding 12 recording sessions and a total of 1800 trials. Thus, about 20% of the samples in the data set correspond to target signals. This protocol is summarized in Figure 2(a) and the directions of the virtual sound sources are shown in Figure 2(b). The neural activity was recorded using a digital electroencephalograph (Active Two, Biosemi, Amsterdam, The Netherlands) with 64 electrodes attached to the subject's scalp using a cap. The electrodes were placed in accordance

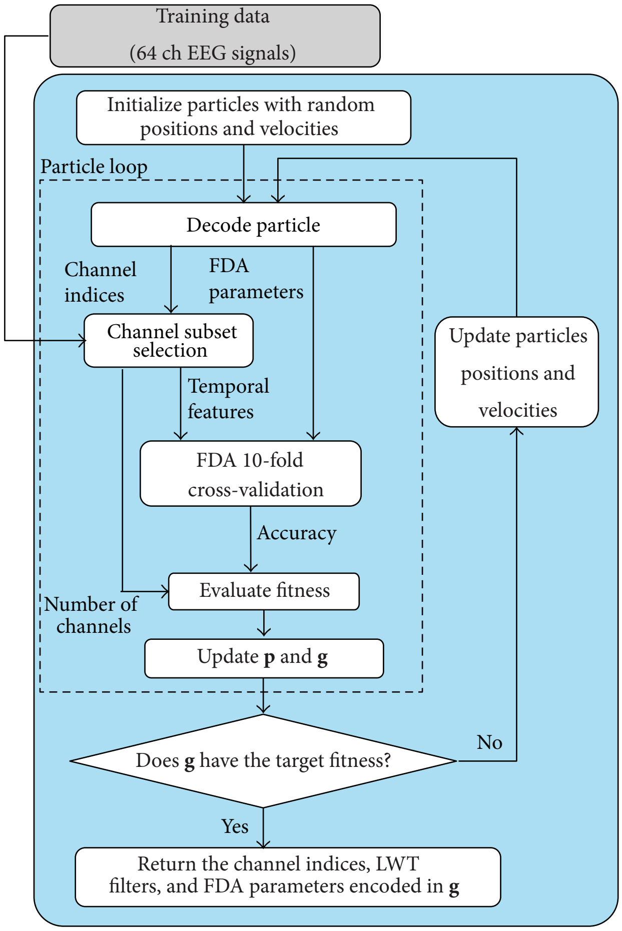

Training data
(64 ch EEG signals)
Initialize particles with random positions and velocities
Particle loop
Decode particle
Channel indices
FDA parameters
Channel subset selection
Temporal features
FDA 10-fold cross-validation
Accuracy
Evaluate fitness
Number of channels
Update **p** and **g**
Update particles positions and velocities
Does **g** have the target fitness?
No
Yes
Return the channel indices, LWT filters, and FDA parameters encoded in **g**

Figure 1: Flowchart of the proposed algorithm.

with the 10–20 system shown in Figure 2(c) and the reference was attached to the earlobes. Twelve healthy men (aged 22–24 years) participated in the experiment.

3.2. PSO Algorithm Setup. Preliminary tests were carried out to choose the best PSO parameters among the values suggested in [20, 21]. The chosen values are summarized as follows. All the PSO simulations were performed using 30 particles until a fitness value equal to 1 was achieved or until 100 iterations were exceeded. For the real part, the inertia parameter w was varied from 0.9 to 0.4 linearly across 100 iterations and had a constant value equal to 1 for the binary part. Both exploration and exploitation constants were set to $c_1 = c_2 = 2$. The search space of the only real component was constrained to $[-1, 1]$. The algorithm employed an invisible wall boundary condition and particles that encoded an empty channel set (i.e., all binary components equal to zero) were deemed as out-of-range. The maximum and minimum velocities were set to −0.1 and 0.1, respectively, for the real part, and −6 and 6, respectively, for the binary part. Regarding the particle's initial conditions at the start of the PSO algorithm, the real component was initialized to a random value in $[-1, 1]$ and each binary component was randomly initialized to either 0 or 1 with equal probability; thus, on average, each particle began the search process with half of the total channels selected.

3.3. Data Preprocessing and Features Used for Classification. Prior to classification, the data was divided in trials, with each trial being the 1100 ms segment of signal starting from −100 ms before the stimulus onset. Then, a zero-phase 3rd order Butterworth band-pass filter with cutoff frequencies of 0.1 Hz and 8 Hz was applied to all the signals across all channels. A baseline correction was performed by subtracting to each trial, at every channel, the mean of the signal in the prestimulus interval $[-100, 0]$ ms. Lastly, in order to reduce the number of temporal samples, each trial was downsampled by taking the average of every 10 samples.

The FDA classifier was fed with 25 temporal samples per EEG channel corresponding to the $[0, 1000]$ ms trial segment. Thus the number of total features ranges from 25 to 1600 depending on the number of channels selected by the PSO algorithm.

3.4. Validation. The performance of the proposed algorithm was assessed by training and testing the algorithm on two disjoint training and test subsets of the original dataset (1800 single-trial samples in total). First, using the training dataset, the algorithm searches for the combination of channels and FDA parameters that maximize the classification accuracy and, when it finishes, it outputs the best configuration that could be found. A final classifier is then trained using the training dataset and the channel and classifier configuration specified by the output of the algorithm. Lastly, the performance of the final classifier is assessed by evaluating its classification accuracy on the test dataset. For the sake of consistency, all the classification accuracies reported in this work are derived from the same criteria used in the fitness function of the MHPSO algorithm, that is, the geometric mean of the target and nontarget accuracies. The subsets were made in such a way that trials of each type (i.e., target and nontarget), each direction (1 to 6), and each session (1 and 2) were present in the subsets with the same proportion as in the original set. Both the training and test sets had approximately 900 samples.

In addition to the single-trial accuracy, in all cases, the averaged-trial accuracy was also assessed. Trial-averaging is a technique usually employed to counteract the high level of noise typically found in EEG signals, and thus improve the classification accuracy. The improved accuracy comes at the cost of a reduced communication speed since several trials must be measured to produce a single estimation. In this work, a M averaged-trial was fabricated from the average classification score (using the score defined in (9)) of M single-trials of the same class randomly chosen with resampling. A set of averaged-trials is built by repeating this process until a given number of trials are fabricated. The number of samples in the averaged-trial dataset was made equal to the number of samples in the single-trial data set (around 900 samples) with the same ratio of target to nontarget samples. In reality, we made a master list containing

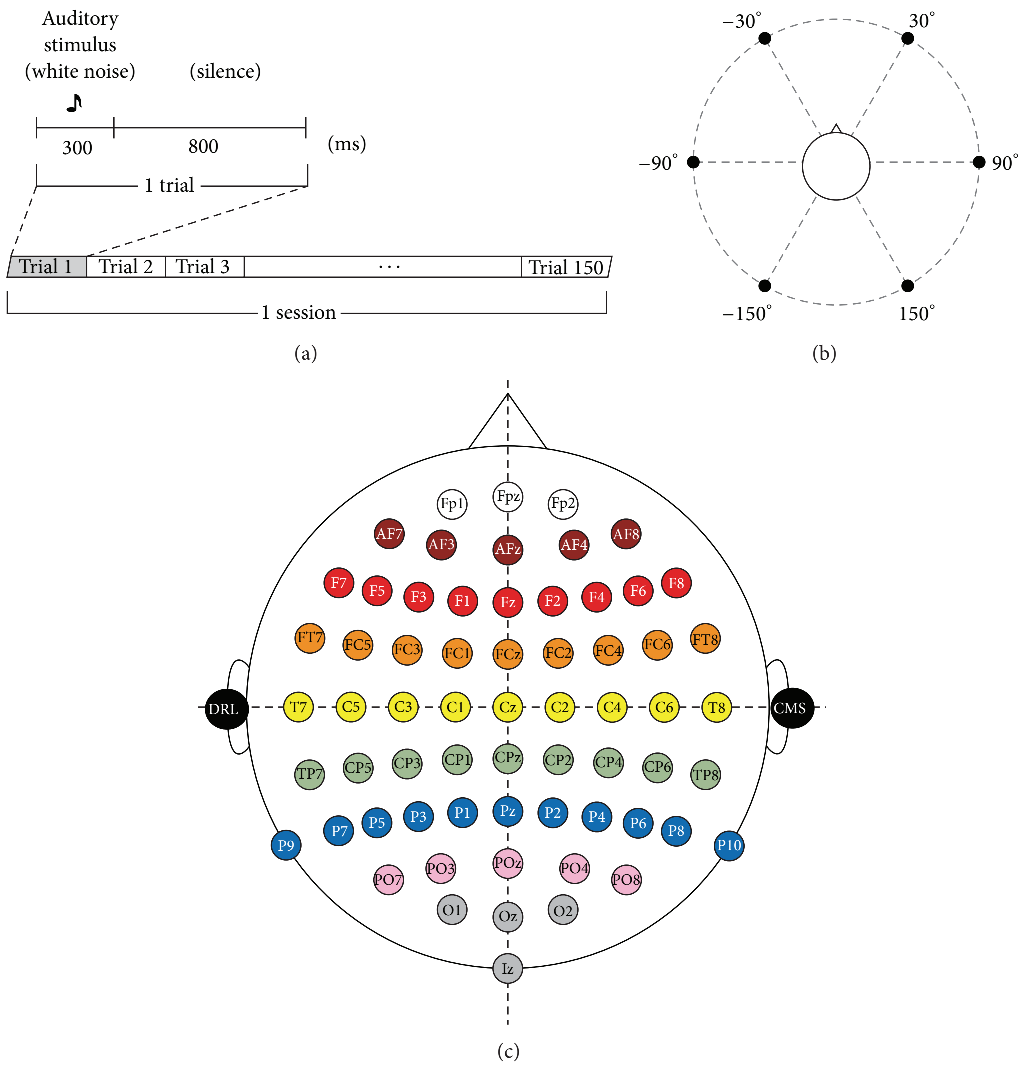


Figure 2: (a) Experimental protocol used to measure the auditory P300 signals. (b) Direction of the virtual sound sources. (c) 64 EEG channel layout used in the experiments. DRL and CMS are the references attached to the earlobes.

the indexes of the single-trials used to make each averaged-trial and used this list in all cases and all subjects to enforce that the averaged-trial datasets were fabricated using exactly the same information and thus ensure a fair comparison between cases. The averaged-trial accuracy was assessed for $M = 2 \ldots 10$.

Lastly, it is important to remark that score averaging was chosen over signal averaging because, if signals are averaged prior to classification, then (1) the number of samples available to train the classifier is reduced, and (2) the P300 component may cancel out if inter-trial jitter is present.

3.5. Simulation Cases. For every subject, eight cases were simulated to study the effect of the fitness function weights on the classification accuracy and number of selected channels that are ultimately achieved by the MHPSO algorithm. These cases are listed below.

Case 1: $w_1 = 1.00$, $w_1 = 0.00$.

Case 2: $w_1 = 0.95$, $w_1 = 0.05$.

Case 3: $w_1 = 0.90$, $w_1 = 0.10$.

Case 4: $w_1 = 0.85$, $w_1 = 0.15$.

Case 5: $w_1 = 0.75$, $w_1 = 0.25$.

Case 6: $w_1 = 0.65$, $w_1 = 0.35$.

Case 7: $w_1 = 0.50$, $w_1 = 0.50$.

Case 8: $w_1 = 0.35$, $w_1 = 0.65$.

Additionally, the algorithm was compared to a version that uses all 64 channels without channel selection and only tunes the FDA parameter (Fixed 64), SWLDA, and a combination of spatial PCA and SWLDA (PCA-SWLDA) similar to the one used in [6]. In the latter cases, the SWLDA algorithm was configured with a feature insertion P value of 0.1 and a feature removal P value of 0.15 as recommended by [7]. In the PCA-SWLDA case, the spatial principal components that accounted for 90% of the variability (the specific number of spatial factors varied between subjects) were used to transform the data prior to classification with SWLDA. In

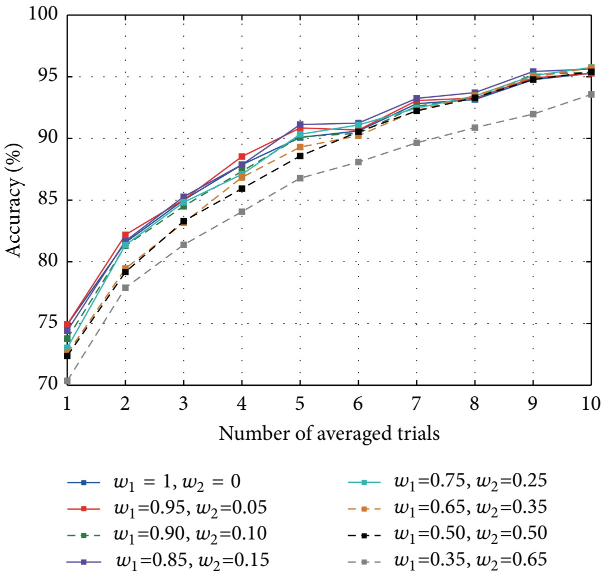


Figure 3: Average classification accuracy of 12 subjects as a function of the number of averaged samples for each case.

the SWLDA case, there were 1600 features to choose from (25 temporal features per channel) and, in the PCA-SWLDA case, there were between 75 and 200 features available for selection (25 transformed temporal features per principal component).

4. Results and Discussion

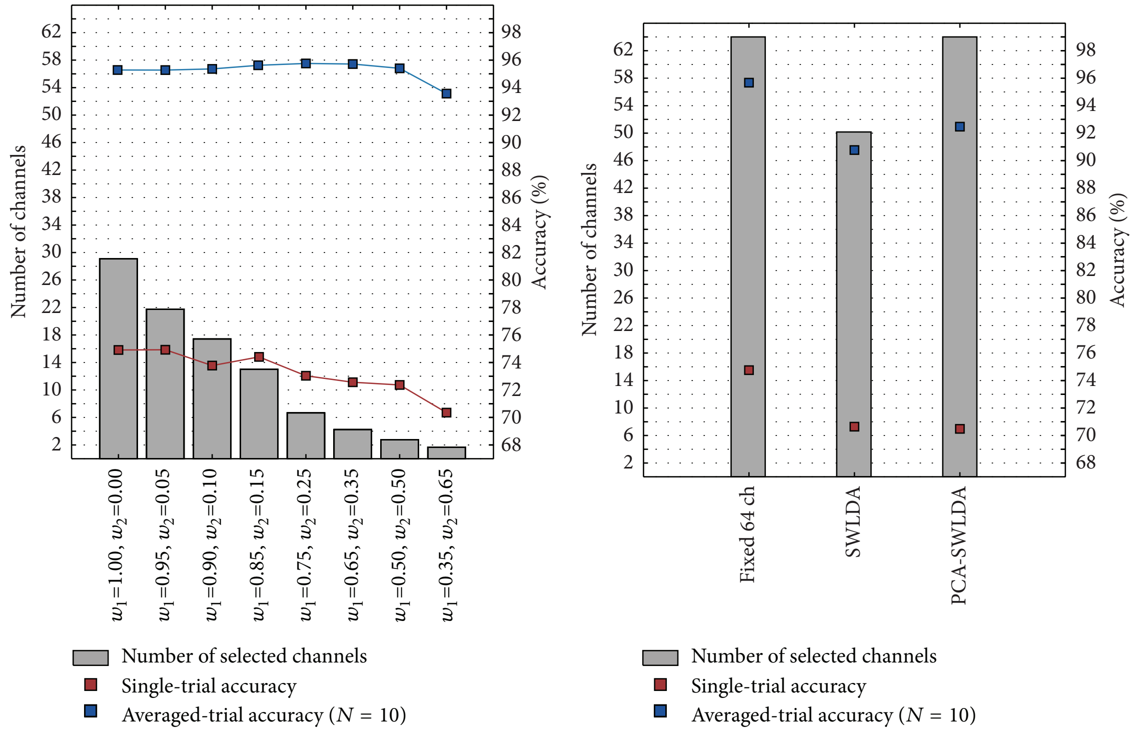
The average classification accuracy of 12 subjects as a function of the number of averaged samples for each MHPSO case is shown in Figure 3. From here, first, an improvement of classification with increasing trial averaging can be seen, as is usually expected. A two-way repeated measures analysis of variance (ANOVA) on the classification accuracy (factors: case and number of averaged samples) revealed a statistically significant difference between cases ($F(7, 77) = 3.440$, $P < 0.005$) and between the number of averaged samples ($F(9, 99) = 266.93$, $P < 0.001$). Since a significant interaction between the case and the number of averaged samples was also found ($F(63, 693) = 2.733$, $P < 0.001$), we performed analysis separately for the single-trial and the 10 averaged-trial cases.

The number of channels selected and the single-trial and 10 averaged-trial accuracies achieved in each MHPSO case are shown in Figure 4(a). The corresponding results for the Fixed 64, SWLDA, and PCA-SWLDA cases are shown in Figure 4(b). This figure illustrates the effect that fitness weights have on the number of selected channels. A one-way repeated measures ANOVA with Greenhouse-Geisser (GG) correction was conducted on the number of channels (factor: case), revealing a significant difference between the number of channels selected by each case ($\varepsilon = 0.593$, $F(4.15, 45.62) = 262.470$, $P < 0.001$). A *post hoc* multiple comparison test based on Holm's method revealed a statistically significant difference between all pairs of cases. Higher w_2/w_1 ratios indeed produced a more aggressive channel selection and selected sets with fewer channels.

While greater values of w_2/w_1 yielded fewer selected channels on average, the single-trial and 10 averaged-trial accuracies did not significantly change. A one-way repeated measures ANOVA with GG correction conducted on the single-trial accuracies (factor: case) found a significant difference between cases ($\varepsilon = 0.434$, $F(3.035, 33.383) = 8.085$, $P < 0.001$), but a *post hoc* Holm test found that these differences were only significant between pairs 1–8 ($P = 0.027$), 2–8 ($P = 0.027$) and 4–8 ($P = 0.027$). A similar analysis performed on the 10 averaged-trial accuracies did not find any significant difference between cases. These results suggest that the number of channels can be reduced without significantly hindering classification accuracy. For example, in Case 7, it can be seen that, with as few as 3 channels, the proposed algorithm could attain a single-trial accuracy slightly slower (around 3%) and an averaged-trial accuracy similar to what would be obtained using the full channel set.

The number of channels selected by the proposed and SWLDA algorithms were also compared. The frequency with which each channel was selected in both cases is shown in Figure 5. Two observations can be made from this figure. The first observation is that the SWLDA algorithm tended to choose most of the channels in most of the subjects, with the least frequently selected channels being chosen at least half of the time. A one-way repeated measures ANOVA conducted on the number of channels with GG correction (factor: case) showed a significant difference between the channels selected by the proposed method and SWLDA ($\varepsilon = 0.344$, $F(2.755, 30.305) = 366.49$, $P < 0.001$). All cases were significantly different to SWLDA ($P < 0.001$ for all pairs), confirming that indeed the proposed algorithm adapts sets with fewer channels. We hypothesized that the random initialization of the particles' binary bits gave the proposed algorithm an unfair advantage over other methods, but simulations where all the particle's binary part were initialized to all ones (i.e., all the channels selected) yielded similar results (results omitted for brevity). The second observation is that the proposed algorithm had a tendency to choose channels over the parietal, frontal, and frontal polar regions, as evidenced by the warm-colored spots over channels Pz, FC1, FCz, FC2, and FPz. While the parietal and frontal channels are in line with the spatial behavior of the P300 [22], the FPz channel is presumably being selected by the proposed algorithm to provide a mean of indirect noise reduction in the FDA classifier as suggested by [5, 23].

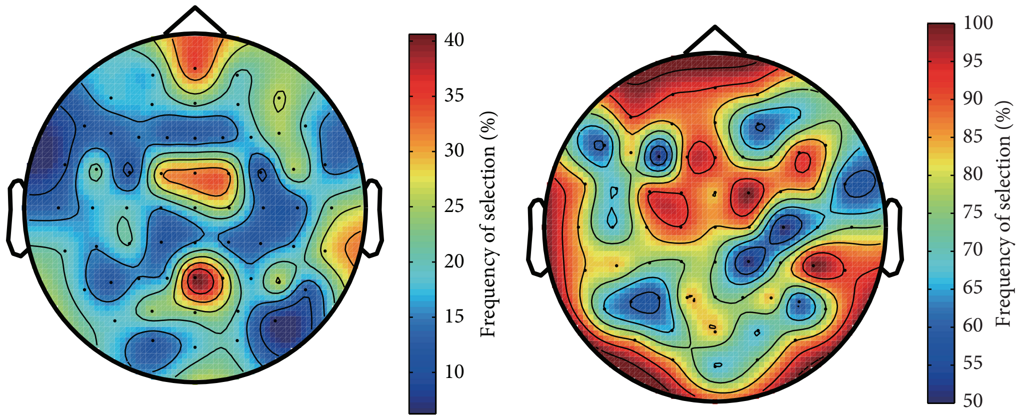
Another way to assess the trade-off relationship between the number of channels and the classification accuracy is to study the Pareto front encountered by the particles throughout the course of the optimization process. The Pareto front in this case represents the boundary at which an improvement of classification accuracy necessarily induces an increase of the number of channels or, conversely, an attempt to reduce the number of channels produces a loss of accuracy. An example of the Pareto front found in each case for a representative subject is shown in Figure 6 to illustrate this phenomenon. By looking at Figure 6, it becomes clear that, for increasing values of w_2/w_1, the positions visited by the swarm gradually



(a)

(b)

Figure 4: Number of selected channels (gray bars), single-trial (red line), and 10 averaged-trial (blue line) classification accuracy for (a) each MHPSO case and (b) the Fixed 64, SWLDA, and PCA-SWLDA cases.



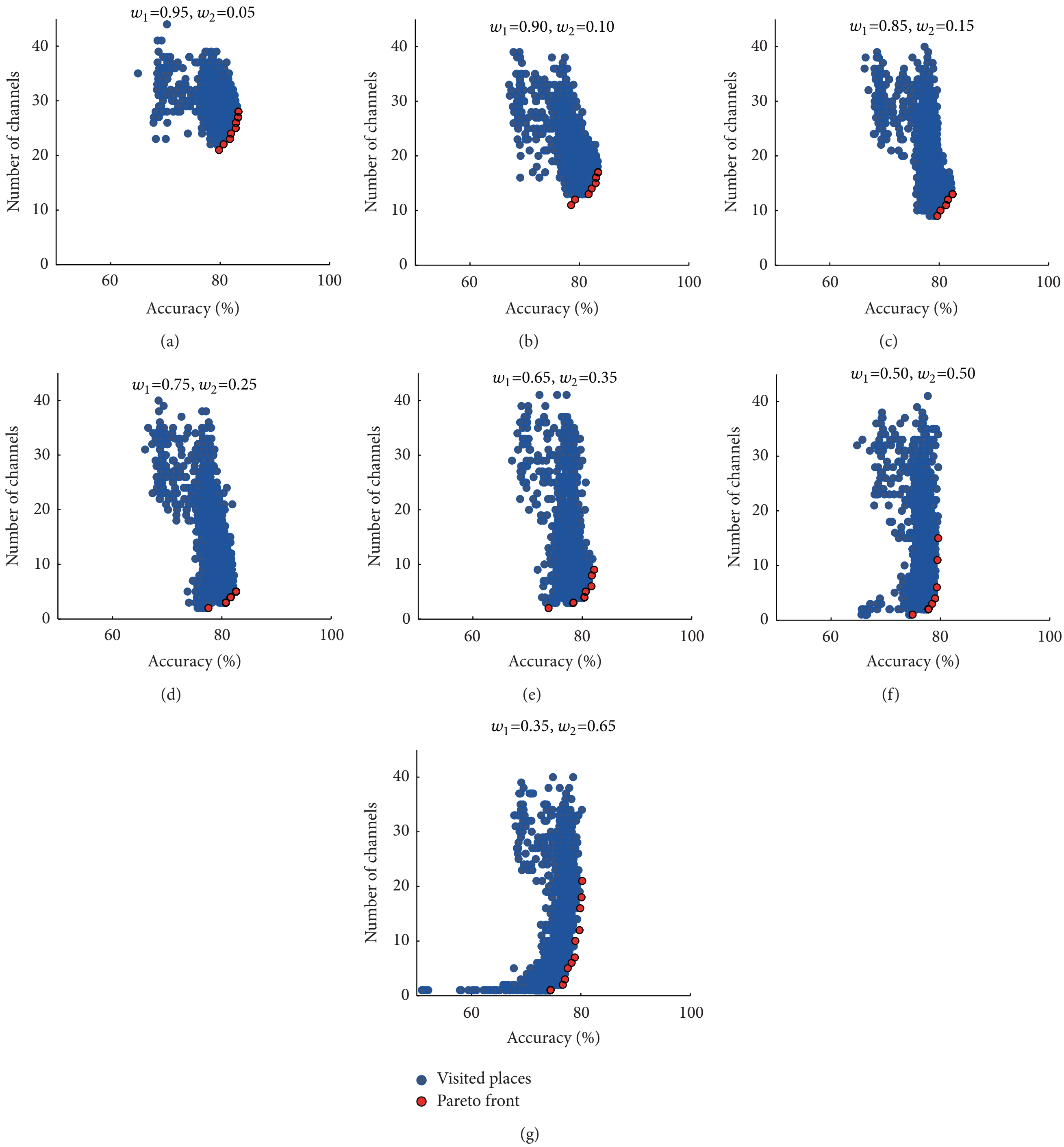
(a)

(b)

Figure 5: Frequency of selection of channels in (a) the proposed algorithm (across all MHPSO cases and subjects) and (b) SWLDA (across all subjects). Note that the color gradations denote different ranges in each figure.

move towards the bottom and that once an accuracy of about 80% is reached, the Pareto front is encountered and a trade-off between number of channels and accuracy begins. Nevertheless, the existence of positions with equal accuracy but different number of channels that can be seen in any of the cases is in line with the abovementioned results.

Lastly, following the results shown in Figure 5(a), it is worth asking if one could use MHPSO to find the most important channels for classification, make a fixed a channel set with a few of those channels, and use it as a general purpose set for every subject. To assess this, we chose the 5 most frequently selected channels across all subjects and cases and trained an algorithm on this reduced set of channels (like Fixed 64, this case only tuned the FDA regularization parameter). We call this case Fixed 5 and compared it to the Fixed 64, SWLDA, PCA-SWLDA, and MHPSO(4) cases. The channel set is shown in Figure 7 and the results are shown in Figure 8. A one-way repeated measures ANOVA



(a) (b) (c) (d) (e) (f) (g)

Figure 6: Illustration of the Pareto front found by the algorithm in each MHPSO case for a representative subject. The vertical and horizontal axes denote the number of selected channels and the average classification accuracy, respectively. The blue dots show the fitness of all the positions visited by the swarm. The red dots represent are the positions that belong to the Pareto front. The first MHPSO case was omitted because the Pareto front in this case is meaningless.

with GG correction conducted on the single-trial accuracies of the 5 algorithms (factor: case) found a significant difference between cases ($\varepsilon = 0.610$, $F(2.440, 26.838) = 9.880$, $P < 0.001$). After a *post hoc* Holm test, it was found that these differences were significant between pairs Fixed 64-SWLDA ($P < 0.001$) and MHPSO(4)-SWLDA ($P < 0.001$). Other significantly different pairs are Fixed 64-PCA-SWLDA ($P = 0.012$), Fixed 64-Fixed 5 ($P = 0.032$), and MHPSO(4)-PCA-SWLDA ($P = 0.029$). No significant differences were found in the 10 averaged-trial case. These results show that, although

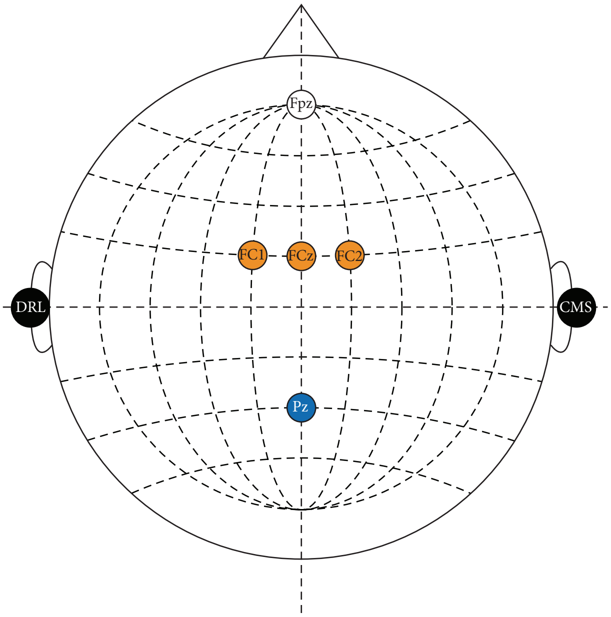


FIGURE 7: Channel set of 5 fixed channels built from the channels that were most frequently selected by the proposed algorithm.

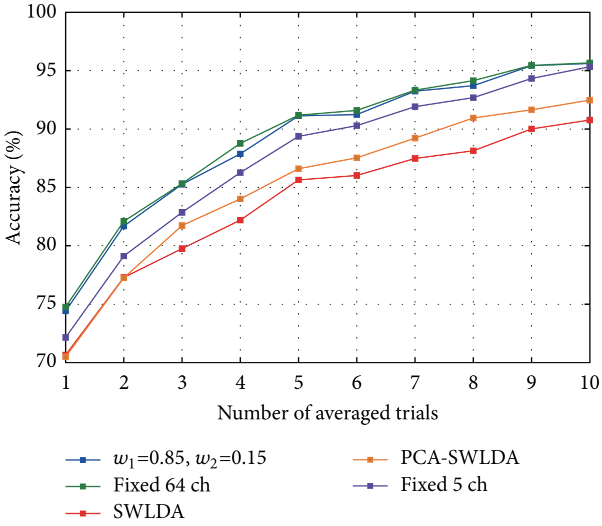


FIGURE 8: Average classification accuracy across 12 subjects achieved by the Fixed 5 case (purple line). Other colors show the accuracy obtained by the Fixed 64, SWLDA, PCA-SWLDA, and the MHPSO(4) cases.

the algorithm using the full channel set provided the best accuracy, the Fixed 5 and MHPSO(4) algorithms yielded a performance comparable to or better than SWLDA and PCA-SWLDA using considerably fewer channels, thus being adequate for BMI systems where portability and simplicity are important. Although there was no significant difference between the Fixed 5 and MHPSO(4) cases (P = 0.197), there also seems to be a difference between using a fixed channel set assessed with MHPSO across all users or using a channel set adapted to each user. This difference, however, may become significant if the number of subjects is increased. Ultimately, the decision of whether to give priority to the accuracy or the compactness of the EEG channel set will depend on the particular constraints of each application. In the case that classification accuracy can be spared, the proposed algorithm can provide hints as to which channels can be used as general purpose set or to adapt a channel set to each particular user.

5. Conclusions

An algorithm based on FDA and MHPSO for the classification of P300 ERP signals was presented. The algorithm used MHPSO to find the FDA parameters and set of the fewest EEG channels that maximized the classification accuracy. The algorithm's performance was evaluated through offline analyses on datasets of auditory ERPs from sound discrimination experiments. The proposed method achieved a higher classification accuracy than that achieved by traditional methods while also using fewer channels. It was also found that it is possible to reduce the number of channels necessary for classification without greatly compromising the classification accuracy. Future work will be aimed at finding why, in addition to channels on the parietal and frontal regions typically associated to the P300 ERP, channels on the frontal polar region were selected. Also, given that the proposed algorithm can be easily extended to spatiotemporal feature selection, further research will focus on a version that tunes more binary variables to select the channel set and the time intervals within each channel that maximize performance. Lastly, future research will also explore the application of other swarm intelligence techniques such as Ant Colony Optimization and the Firefly Algorithm and compare it with the results obtained with MHPSO.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by JSPS KAKENHI Grant nos. 24300051, 24650104, and 24800027.

References

[1] M. Kogure, M. Ebisawa, S. Yano, S. Matsuzaki, and Y. Wada, "Estimation of human's chosen direction using auditory-evoked event-related potentials," Tech. Rep., IEICE, Tokyo, Japan, 2011.

[2] M. Ebisawa, M. Kogure, S.-H. Yano, S.-I. Matsuzaki, and Y. Wada, "Estimation of direction of attention using EEG and out-of-head sound localization," in *Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS '11)*, vol. 2011, pp. 7417–7420, Boston, Mass, USA, September 2011.

[3] I. Nambu, M. Ebisawa, M. Kogure, S. Yano, H. Hokari, and Y. Wada, "Estimating the intended sound direction of the user: toward an auditory brain-computer interface using out-of-head sound localization," *PloS ONE*, vol. 8, no. 2, article e57174, 2013.

[4] D. J. Krusienski, E. W. Sellers, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "Toward enhanced P300 speller performance," *Journal of Neuroscience Methods*, vol. 167, no. 1, pp. 15–21, 2008.

[5] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components—a tutorial," *NeuroImage*, vol. 56, no. 2, pp. 814–825, 2011.

[6] S. Kamp, A. R. Murphy, and E. Donchin, "The component structure of event-related potentials in the P300 speller paradigm," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 21, no. 6, pp. 897–907, 2013.

[7] D. J. Krusienski, E. W. Sellers, F. Cabestaing et al., "A comparison of classification techniques for the P300 Speller," *Journal of Neural Engineering*, vol. 3, no. 4, article 007, pp. 299–305, 2006.

[8] S. Derksen and H. J. Keselman, "Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables," *British Journal of Mathematical and Statistical Psychology*, vol. 45, no. 2, pp. 265–282, 1992.

[9] A. Rakotomamonjy and V. Guigue, "BCI competition III: dataset II- ensemble of SVMs for BCI P300 speller," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 3, pp. 1147–1154, 2008.

[10] B. Blankertz, "BCI Competition III Webpage," 2004.

[11] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 41–56, 2008.

[12] D. J. Krusienski, E. W. Sellers, and T. M. Vaughan, "Common spatio-temporal patterns for the P300 speller," in *Proceedings of the 3rd International IEEE EMBS Conference on Neural Engineering (CNE '07)*, pp. 421–424, Kohala Coast, Hawaii, USA, May 2007.

[13] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, vol. 4, pp. 1942–1948, Perth, Australia, December 1995.

[14] R. Poli, "Analysis of the publications on the applications of particle swarm optimisation," *Journal of Artificial Evolution and Applications*, vol. 2008, Article ID 685175, 10 pages, 2008.

[15] N. Jin and Y. Rahmat-Samii, "Hybrid real-binary particle swarm optimization (HPSO) in engineering electromagnetics," *IEEE Transactions on Antennas and Propagation*, vol. 58, no. 12, pp. 3786–3794, 2010.

[16] J. Kennedy and R. Eberhart, "A discrete binary version of the particle swarm algorithm," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, vol. 5, pp. 4104–4108, Orlando, Fla, USA, October 1997.

[17] N. Jin and Y. Rahmat-Samii, "Advances in particle swarm optimization for antenna designs: real-number, binary, single-objective and multiobjective implementations," *IEEE Transactions on Antennas and Propagation*, vol. 55, no. 3, part 1, pp. 556–567, 2007.

[18] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 7, pp. 179–188, 1936.

[19] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, New York, NY, USA, 2004.

[20] R. Eberhart and Y. Shi, "Particle swarm optimization: developments, applications and resources," in *Proceedings of the Congress on Evolutionary Computation*, vol. 1, pp. 81–86, Seoul, Republic of Korea, May 2001.

[21] J. Robinson and Y. Rahmat-Samii, "Particle swarm optimization in electromagnetics," *IEEE Transactions on Antennas and Propagation*, vol. 52, no. 2, pp. 397–407, 2004.

[22] J. Polich, "Updating P300: an integrative theory of P3a and P3b," *Clinical Neurophysiology*, vol. 118, no. 10, pp. 2128–2148, 2007.

[23] D. McFarland, C. Anderson, K. R. Müller, A. Schlögl, and D. Krusienski, "BCI meeting 2005-workshop on BCI signal processing: feature extraction and translation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 135–138, 2006.