INTERNATIONAL JOURNAL OF ADAPTIVE CONTROL AND SIGNAL PROCESSING
Int. J. Adapt. Control Signal Process. 2012; **26**:482–495
Published online 23 December 2011 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/acs.1293

Unsupervised adaptation of electroencephalogram signal processing based on fuzzy C-means algorithm

Guangquan Liu, Dingguo Zhang*,†, Jianjun Meng, Gan Huang and Xiangyang Zhu

State Key Laboratory of Mechanical System and Vibration, Shanghai Jiao Tong University , Shanghai, 200240, China

SUMMARY

This paper studies an unsupervised approach for online adaptation of electroencephalogram (EEG) based brain–computer interface (BCI). The approach is based on the fuzzy C-means (FCM) algorithm. It can be used to improve the adaptability of BCIs to the change in brain states by online updating the linear discriminant analysis classifier. In order to evaluate the performance of the proposed approach, we applied it to a set of simulation data and compared with other unsupervised adaptation algorithms. The results show that the FCM-based algorithm can achieve a desirable capability in adapting to changes and discovering class information from unlabeled data. The algorithm has also been tested by the real EEG data recorded in experiments in our laboratory and the data from other sources (set IIb of the BCI Competition IV). The results of real data are consistent with that of simulation data. Copyright © 2011 John Wiley & Sons, Ltd.

Received 27 April 2011; Revised 17 August 2011; Accepted 1 November 2011

KEY WORDS: electroencephalogram (EEG); brain–computer interface; unsupervised adaptation; fuzzy C-means (FCM)

1. INTRODUCTION

As an emerging technique, brain–computer interfacing provides a versatile tool for people to interact environment or control mechanical devices directly by intent rather than by neuromuscular pathway [1]. Brain–computer interfaces (BCIs) have potential use in a wide range of areas. For instance, they may provide disabled people suffering from neuromuscular disorders with alternative ways for communication and control. In practice, the brain signals such as electroencephalogram (EEG), which encode the information of human intention, usually play an important role in the development of BCIs. As a well-recognized fact [2–5], EEG signals are usually nonstationary in nature. The nonstationarity may be due to various reasons, for example, ambient noise, changes in the concentration or excitation level, changes in the mental task involved, changes in the impedance or even position of the electrodes, demands for visual processing, and influence of feedback, fatigue, and artifacts such as swallowing and blinking [2–13]. Because of nonstationarity, a BCI system trained on one session may become ill-suited to the succeeding sessions.

In order to improve the adaptability of BCIs, researchers have extensively investigated adaptation algorithms in the literature. In [14–16], the classifier was updated manually after several runs to adapt the change in the brain states, according to the experimenter's experience and judgment. In [2–4, 13, 17–20], the classifier was updated automatically by different machine learning methods. Most of these methods need true label information to update the classifier. In other words, these methods are supervised.

*Correspondence to: Dingguo Zhang, State Key Laboratory of Mechanical System and Vibration, Shanghai Jiao Tong University, Shanghai, 200240, China.

†E-mail: dgzhang@sjtu.edu.cn

However, in a practical application scenario, the real intention of the subject may be unknown to the system [2, 6, 11]. For this reason, some research groups have tried unsupervised adaptation algorithms. In [21], the unsupervised inter-subject validation of a P300-based BCI is studied. For EEG-based BCIs, most researches focus on the adaptation along time for a fixed subject. Three types of unsupervised adaptation procedures for the linear discriminant analysis (LDA) classifier were proposed and analyzed in [11]. These procedures are based on an assumption that during the online sessions, the vector connecting the two class means keeps nearly constant. Without such assumptions, Gan [22] introduced an incremental adaptation procedure for LDA and Bayesian classifiers, which used the estimated label to guide the updating of the classifier. This is a 'decision-directed' approach, which may have some potential disadvantages. If a trial is wrongly classified, the classifier will be misled by it, although this error can be reduced by finely tuning of the update coefficient. To avoid such disadvantages, some non-incremental unsupervised clustering methods have also been reported. In [23], Eren *et al*. used Gaussian mixture model (GMM) for unsupervised clustering of EEG patterns, but this work did not check the online performance. Recently, Hasan and Gan studied the online performance of the GMM method in [24]. In [25], Blumberg *et al*. proposed an adaptive LDA (ALDA) method for simulated online clustering of EEG patterns. This method is actually a GMM-based adaptation method, although not explicitly named. This method has two limits, namely, the fixed initial parameter set and the need to estimate the covariance matrices.

The first problem is solved by initializing the model parameters of the GMM method [26]. This improved GMM method shows a better performance than the basic GMM. However, the second problem still exists. In order to solve this problem, we adopt the fuzzy C-means (FCM) algorithm for the unsupervised adaptation of the LDA classifier in a simulated online BCI scenario. The FCM algorithm has been used in many fields, for example, medical image segmentation [27], unsupervised quantification of epileptic EEG patterns [28], and classification of wrist pulse signal patterns [29], but according to our knowledge, it has not been used in the unsupervised adaptation of BCIs. The preliminary results of this paper were reported in [30]. This method has some advantages. Firstly, the classifier is not updated incrementally, but recomputed from the recent unlabeled trials every time a new trial becomes available, which may reduce the potential disadvantage of error accumulation to some extent. Secondly, during the iterative parameter estimation, only the class means need to be estimated, whereas the covariance matrices can be estimated at the last iteration only once. Thirdly, we treat the initial parameter set (two class means) used in the FCM as variational and update it along time according to the historical estimations. This makes the method more adaptive to significant changes in the data distribution during long-term use. In order to investigate its performance in different situations, we applied the proposed method to several types of constructed artificial data with different properties, as well as real EEG data collected from 22 experiments in our laboratory and other resources (data set IIb of the BCI Competition IV). The performance of the FCM is compared with a static LDA that is not updated, a supervised adaptive LDA, an incrementally updated LDA, a GMM-based adaptive LDA, and a common mean change based adaptive LDA. Our method shows a better performance than the other unsupervised adaptation methods in most cases. We also investigate how the data properties (separability and nonstationarity) affect the performances of different adaptation methods. Analysis on artificial data shows that our method is quite adaptive to nonstationarity and can effectively discover the class information from unlabeled data. Results of the real data show a similar performance to the artificial data.

The paper is organized as follows: Section 2 describes the methods, including the feature extractor and classifier, the proposed FCM-based method, and other existing adaptation algorithms. Section 3 describes the constructed artificial data, our experimental setup, and the BCI competition data; Section 4 reports the results and does some discussion. Section 5 concludes the paper.

2. METHODS

2.1. Feature extraction

Common spatial patterns (CSP) is used as a feature extractor in this paper. CSP is a supervised spatial filtering method for two-class discrimination problems, which finds directions that maximize the variance for one class and at the same time minimize the variance for the other class. Mathematically, CSP is realized by simultaneous diagonalization of the covariance matrices for the two classes [31]. The formulas of this algorithm can be found in [32]. The normalized log-variances of six most discriminative components are used as features. The transformation to logarithmic values makes the distribution of the features approximately normal.

2.2. LDA classifier

Fisher's LDA [33] is used as a classifier in this paper. If μ_1 and μ_2 are the means of the two classes, and Σ_1 and Σ_2 are the corresponding covariance matrices, then an LDA classifier can be determined by these parameters. The weight of the classifier is

$$w = (\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2) \tag{1}$$

and the bias is

$$b = -w^{\mathrm{T}}\left(\frac{\mu_1 + \mu_2}{2}\right) \tag{2}$$

where w^{T} means transpose of vector w. For each feature vector x, the classifier output is

$$y = w^{\mathrm{T}}x + b \tag{3}$$

If $y > 0$, then x is classified into class 1; otherwise, class 2.

2.3. Supervised adaptation of classifier

We denote the parameter set as $\Theta = \{\mu_1, \mu_2, \Sigma_1, \Sigma_2\}$. As the LDA classifier can be fully determined by Θ, in order to update the classifier, the main work we need to do is to update Θ.

In a supervised scenario, where all the trials before the current time are provided with a true label, the adaptation work is relatively easier. If x_k is the latest feature vector whose true label is already known, we update Θ in the following way:

$$\mu_i(k) = \mu_i(k-1)(1-UC) + x_k \cdot UC \tag{4}$$

$$\begin{aligned}\Sigma_i(k) = {} & \Sigma_i(k-1)(1-UC) \\ & + (x_k - \mu_i(k))(x_k - \mu_i(k))^{\mathrm{T}} \cdot UC\end{aligned} \tag{5}$$

where i is the true label of x_k and UC is the update coefficient.

2.4. Unsupervised adaptation of classifier

As mentioned in [4], 'in a realistic BCI scenario, the labels of ongoing trials may not always be available'. In such a case, the BCI system has to be updated in an unsupervised manner, if adaptation is necessary. The FCM method [34] has been used for many unsupervised clustering problems; however, it has not been applied to the adaptation of BCIs. In this work, we adopt this method for the unsupervised adaptation of the LDA classifier. The adaptation of the feature extractor is another important topic, which is not considered in the present work.

2.4.1. Fuzzy C-means. In a clustering problem, fuzzy clustering algorithms partition the feature space into overlapping areas (classes) on the basis of the similarity among feature vectors, and each feature vector is provided with a membership value to each class. However, in this work, the FCM is not used directly as a clustering method, but used as a technique for the unsupervised estimation of the parameter set Θ.

The FCM is an iterative algorithm, and in each iteration there are two steps.

Step 1
Given the class means, the Euclidean distance between the kth feature vector and the ith class mean can be computed:

$$d_{ik} = \sqrt{(x_k - \mu_i^{(t)})^{\mathrm{T}}(x_k - \mu_i^{(t)})} \tag{6}$$

where t is the iteration index and k is the trial index, the range of which will be discussed later. Then, the membership value can be computed as

$$p_{ik} = \frac{1}{\sum_{j=1}^{C}(d_{ik}^2/d_{jk}^2)^{1/(m-1)}} \tag{7}$$

where m is the fuzziness exponent ($m > 1$) and C is the number of classes (in this work, $C = 2$).

Step 2
Given the membership values, the class means can be updated as

$$\mu_i^{(t+1)} = \frac{\sum_k (p_{ik})^m x_k}{\sum_k (p_{ik})^m} \tag{8}$$

Similarly, the covariance of each class can also be estimated:

$$\Sigma_i = \frac{\sum_k (p_{ik})^m C_{ik}}{\sum_k (p_{ik})^m} \tag{9}$$

where $C_{ik} = (x_k - \mu_i)(x_k - \mu_i)^{\mathrm{T}}$. As Σ_i is not involved in Step 1 of each iteration, it can be computed only once at the last iteration. This not only saves computational time but also reduces the chance of inappropriate estimation of covariance matrices and membership values in the iteration.

Some remarks about this method should be stated.

- The objective function. The iterative procedure of FCM is actually trying to minimize such an objective function

$$f = \sum_{i=1}^{C}\sum_k p_{ik}^m d_{ik} \tag{10}$$

  under constraint

$$\sum_{i=1}^{C} p_{ik} = 1 \quad \text{for all } k \tag{11}$$

  where $p_{ik} \in [0, 1]$.
- The range of k. If I_{cu} is the current trial index, and N_{ut} is the size of the unlabeled 'training' set used to update the classifier, then $k \in [I_{\mathrm{cu}} - N_{\mathrm{ut}} + 1, I_{\mathrm{cu}}]$.
- The initial parameter set Θ_{init}. We treat Θ_{init} as a variational parameter but not a constant. The current Θ_{init} is estimated on the basis of the historical estimations of Θ. Let $\Theta^{(0)}$ denote the Θ computed from the labeled training session, $\Theta_{\mathrm{init}}^{(k)}$ denote the Θ_{init} used for estimation when

the kth unlabeled trial becomes available and let $\Theta^{(k)}$ denote the estimated Θ after the kth unlabeled trial. Then, we determine $\Theta_{\text{init}}^{(k)}$ as

$$\Theta_{\text{init}}^{(k)} = \begin{cases} \Theta^{(0)}, & \text{if } k \leqslant N_{\text{hi}}, \\ \frac{1}{N_{\text{hi}}} \sum_{j=k-N_{\text{hi}}}^{k-1} \Theta^{(j)}, & \text{else} \end{cases} \tag{12}$$

where N_{hi} is the size of the historical set of the estimations. We do not use a constant or random Θ_{init}, in order to avoid the potential danger that the class distribution shifts far away from the initial distribution or even becomes mirrored to the initial distribution (corresponding to a class label switch) during long-term use.

- The number of iterations N_{it}. In general, the iteration stops when the difference of the membership matrix between two successive iterations is smaller than a threshold or the number of iteration reaches a maximal limit. In an online scenario, a trade-off should be made between the estimation precision and the consumed computational time. In our work, we let $N_{\text{it}} = 1$, as we find that one iteration is sufficient for the adaptation, and more iterations cannot further improve the performance.
- The fuzziness exponent m. m is a hyper-parameter that controls the fuzziness degree of the model. If $m \to +\infty$, then the membership value of each feature vector to each class will be $1/C$; if $m \to 1$, then the method will become a hard C-means method, where the membership value is either 1 or 0. We use a typical choice of $m = 2$. A detailed discussion about this choice will be presented in Section 4.

2.4.2. Other existing methods. As references, three existing unsupervised adaptation methods are compared with the FCM method.

An incremental updating method (denoted as INCRE) is proposed to update the mean and covariance of each class in [22]. This method is a ‘decision-directed’ approach [33], which may have some disadvantages. If the prior classifier is not good enough or if several unfortunate trials are encountered, which often happens when the nonstationarity is severe, the updating will make the classifier worse. In this work, the INCRE method is implemented in the following way: step 1, an old LDA (determined by the labeled training trials and the previous unlabeled trials), is used to classify the coming trial, obtaining an estimated label; Step 2, if the absolute value of the LDA output exceeds a predefined threshold, that is, the classification is confident enough, we update the classifier with this trial; otherwise, make no change. The updated new LDA is used as the ‘old’ LDA for the next trial.

A GMM-based unsupervised method for clustering of EEG patterns is introduced in [23], but the online adaptation was not considered. Blumberg *et al.* proposed an ALDA method to update the LDA classifier in a simulated online manner in [25], which is also a kind of GMM-based adaptation method in nature. In this method, the expectation maximization (EM) method was used to estimate the means and a common covariance of the two classes. In the E-step, the probability of each trial belonging to each class is calculated according to the current model parameters (class means and covariances of the Gaussian distributions). In the M-step, given the data’s class membership distribution, the model parameters are re-estimated to maximize the likelihood. This procedure is repeated until it converges or a predefined maximal iteration number is reached. In this work, the iteration number is set to 1. The initial data distributions are estimated from the labeled training session. There are mainly two limits in this method. One limit is that if the feature distribution shifts far away from the training session during long-term use, this initial parameter will not be valid any more. Another limit is that as the clusters are assumed to be Gaussian, the covariance matrices of each class have to be estimated in every iteration of the EM algorithm. However, compared with the mean vector, the covariance matrix is prone to be inappropriately estimated because it has more elements ($l \times l$, with l the dimensionality of the feature space).

In [11], three unsupervised adaptation methods were proposed, namely, a common mean change method (CMean), a common mean and covariance change method (CMean-CCov), and a rotation method (Rotation). It was reported that these three methods performed very similarly, with a slight

DOI: 10.1002/acs

advantage of the CMean classifier. The CMean method assumes that the two class covariances and the vector connecting the two class means are stable, and only the common mean of the two classes changes. Therefore, the LDA classifier can be updated by just adjusting its bias b, with the weight w unchanged. The bias b can be determined by the common mean of the unlabeled trials as $b = -w^{\mathrm{T}}u$. This method is very simple and straightforward, but it is effective if the assumption is valid.

3. APPLICATION

3.1. *Artificial data*

In order to study the performance of the proposed method, and the factors that influence the performance, besides the recorded real EEG data, we also construct groups of artificial data. As these data are artificially constructed, we can control their properties, for example, separability, type of nonstationarity, and degree of nonstationarity. In such a way, we can thoroughly investigate how these properties affect the adaptation of a BCI.

As reported in [4], there are mainly two types of changes in EEG data, namely, shift in the transition from one session to another and gradual change in the course of a single session. It was reported that 'the major detrimental influence on the classification performance is caused by the initial shift from training to the test scenario' [4]. In our work, we construct three types of artificial data. The first type shifts between sessions but keeps stationary within one session (SHIFT). The second type changes gradually in the course of a session without the initial shift (GRAD). The third type contains these two types of changes (BOTH).

As the main purpose of this work is to investigate the adaptation of the classifier, we directly simulate the feature vectors rather than EEG signals. The artificial features are constructed in a procedure as follows: (i) Two classes of static feature vectors are generated randomly, with each class following a multi-dimensional (here six-dimensional) Gaussian distribution. The distance between the two class means is controlled to determine the separability of the data. (ii) The features are divided into three sessions, with the first one as labeled training session and the last two as unlabeled test sessions. (iii) A bias with the same dimension is added to the features. The type of the bias can be SHIFT, GRAD, or BOTH. The magnitude of the bias is controlled to determine the degree of nonstationarity of the data. (iv) Considering that the features are generated randomly, there is a chance of producing abnormal data. In order to reduce the influence of abnormal data, for each separability and nonstationarity level, we generated 10 data sets and their error rates are averaged.

The separability of the data is indicated by a parameter r_{cls}, which is defined as

$$r_{\mathrm{cls}} = \frac{|\mu_1 - \mu_2|\sqrt{n_1 n_2}}{\sqrt{\left|(\Sigma_1 + \Sigma_2) \cdot \frac{(\mu_1-\mu_2)}{|\mu_1-\mu_2|}\right|(n_1 + n_2)}} \tag{13}$$

where $|\cdot|$ means the norm of a vector and n_i denotes the number of trials in class $i \in \{1, 2\}$. r_{cls} measures how well the two classes are separated. A bigger r_{cls} means that the two class means are farther away from each other relative to the variance in this direction. In this paper, we denote r_{cls} to indicate the separability of the two classes, which is different from the following r_{chg}.

The degree of nonstationarity of the data is indicated by r_{chg}, which is defined as

$$r_{\mathrm{chg}} = \frac{|\mu_{\mathrm{tr}} - \mu_{\mathrm{te}}|\sqrt{n_{\mathrm{tr}} n_{\mathrm{te}}}}{\sqrt{\left|(\Sigma_{\mathrm{tr}} + \Sigma_{\mathrm{te}}) \cdot \frac{(\mu_{\mathrm{tr}}-\mu_{\mathrm{te}})}{|\mu_{\mathrm{tr}}-\mu_{\mathrm{te}}|}\right|(n_{\mathrm{tr}} + n_{\mathrm{te}})}} \tag{14}$$

where the subscripts 'tr' and 'te' indicate the labeled training data set and the unlabeled test data set, respectively. This parameter can be interpreted as the difference between the training data and the test data. A bigger r_{chg} means that the test data are more different from the training data, that is, the nonstationarity is severer.

We use these hyper-parameters to control the property of artificial data, so as to investigate how well the methods in this paper can perform in different situations.

DOI: 10.1002/acs

3.2. Experimental setup and SJTU data

3.2.1. Subjects. Eight right-handed healthy subjects (six men and two women, age 23–28 years) took part in the experiment. None of them had an experience of BCI experiment before. The volunteers were paid for their participation.

3.2.2. Procedure. The subjects were seated in a comfortable armchair about 2 m in front of a computer monitor. They were instructed to keep still and avoid blinking during a trial. At the beginning of each trial, the screen was black. One second later, a fixation cross appeared in the center of the screen. Another second later, an arrow pointing to either left or right was added to the cross, indicating the imagination of left-hand or right-hand movement. The arrowed cross was shown until the end of second 5. During this period (from the beginning of second 3 to the end of second 5), the subject had to imagine left or right (corresponding to the cue) hand movements. The type of movement was decided by the subject herself/himself, for example, patting a ball or pulling a brake. At the end of second 5, a feedback of this single trial was provided by moving the arrowed cross to the left or right side of the screen, according to the classifier output. This was the situation in a feedback session. If the session was a training session, then no feedback was provided, that is, the arrowed cross remained at the center of the screen until this trial was over. After a random interval varying from 1.5 to 2.5 s, the next trial began. The sequence of left and right trials was randomized, and the chance for each class was equal. In each run, 10 left and 10 right trials were performed. There were five runs in each session and three sessions in each data set, resulting in 300 trials in each data set. The first session was used as a training session, and the last two sessions were online feedback sessions. There was a 1-min break between two successive runs and a 10-min break between two successive sessions. The break could alleviate the subjects' fatigue and avoid them from becoming tired. A data set was recorded on the same day. Each subject took one to three experiments, resulting in a total of 22 data sets.

In the experiments, the online feedback was provided at the end of each trial but not continuously during the whole trial. The signals used for calculation are the complete event-related portion (from second 3 to second 5). We also managed to provide a continuous feedback in some other experiments. However, the subjects reported that it would be more difficult for them to concentrate while the cursor was moving all the time. The reason may be that for naive subjects who are not well trained, a simpler feedback may be helpful. Although for well-trained subjects, a continuous feedback may excite them to achieve a better performance.

During the online experiments, the classifier actually used to give a feedback was a LDA classifier, which was updated in a supervised manner. All the data were saved for further analysis. The unsupervised adaptation methods were investigated in a simulated online manner. In the simulation scenario, every step was taken as it was performed in the online scenario. Only the data before the current time were used for computing, and the data after the current time were treated as completely unseen. In other words, the procedure was causal.

3.2.3. Recordings. EEG signals were recorded using a SynAmps system (Neuroscan, Charlotte, NC, USA). Signals from 21 channels over central and related motor areas were used for classification. The grounding electrode was mounted on the forehead and reference electrodes on the left and right mastoids. The electrodes were placed according to the extended 10/20-system [35] (see Figure 1). Horizontal and vertical electrooculographs (EOGs) were recorded for the purpose of artifact detection and were not used for classification. The EEGs were first filtered by the recording system in a 5- to 30-Hz frequency band, and the sampling rate was 1000 Hz. Before feature extracting and classifying, the signals were down-sampled to 200 Hz and re-filtered in 8–30 Hz by an FIR filter. By the high-pass filtering, low-frequency EOG artifacts were also removed.

In the rest of this paper, we will refer to these data as SJTU data.

3.3. Graz data

For the purpose of validation and comparison, we also apply the methods to the data set IIb of BCI Competition IV, which is provided by Graz University of Technology, Austria. It consists of EEG

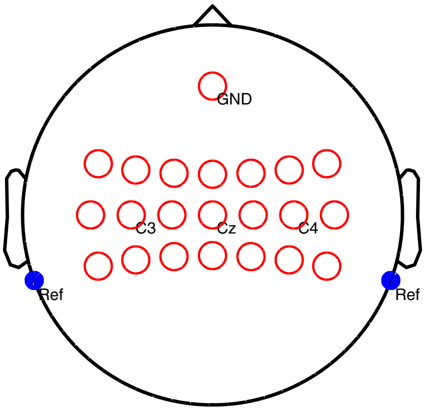


Figure 1. Placement of the 21 EEG electrodes in the SJTU experiments. The three marked electrodes are C3, Cz, and C4 respectively. ‘GND’ means ground electrode, and ‘Ref’ means reference electrodes.

data from nine subjects, with five sessions for each subject. The data are related to two-class motor imagery. Details can be found in [36].

Using CSP applied to signals from different frequency bands and static LDA without adaptation, we took part in the BCI Competition IV, and obtained the second place of data set IIb. In the rest of this paper, we will refer to these data as Graz data. It should be noted that only the last three sessions of each subject are used for computation as the first two sessions have a different experimental paradigm.

4. RESULTS AND DISCUSSION

For both artificial and real data, we check the performance of six methods, namely, static LDA without adaptation (STATIC), supervised adaptation (SUPER), incremental adaptation (INCRE), GMM-based adaptation (GMM), common mean changed based adaptation (CMean), and the proposed FCM-based approach (FCM). Hyper-parameters for each method are optimized on the basis of the training data. The training sessions are further divided into three parts: one training part and two validation parts. A 3×3 cross-validation is performed to find the hyper-parameters, which maximize the averaged accuracy on the validation part. Considering that the validation is based on much fewer trials than the online (simulated) application, some hyper-parameters (such as N_{hi} and N_{ut} in the FCM) are forcedly enlarged.

Figure 2 shows the error rate means and standard deviations of the six methods on the artificial data (SHIFT, GRAD, BOTH), the SJTU data, and the Graz data. The error rates in Figure 2 are averaged over all data sets with the same data type. From this figure, we can see that for all data types, FCM shows a significant improvement over STATIC, INCRE, and GMM. For the artificial data and the SJTU data, FCM has a slight advantage over CMean, but not significant. For the Graz data, CMean achieves the best performance, which is even slightly better than SUPER. However, it is notable that for the Graz data, the differences among these six methods are quite small (within 1.5%). For all data types, INCRE shows a little improvement over STATIC; GMM falls into a level between FCM and INCRE. SUPER has a stable but not great advantage over FCM and CMean for most data types. Results of the three types of artificial data are very similar, implying that the type of nonstationarity does not affect the performance significantly.

4.1. Artificial data

The artificial data have the same framework as the real data, that is, three sessions for each data set, with the first one as a training session and the other two as test sessions. By varying r_{cls} and r_{chg} each among 15 degrees, we obtain 225 groups of data sets, each group with fixed r_{cls} and r_{chg}. In each

DOI: 10.1002/acs

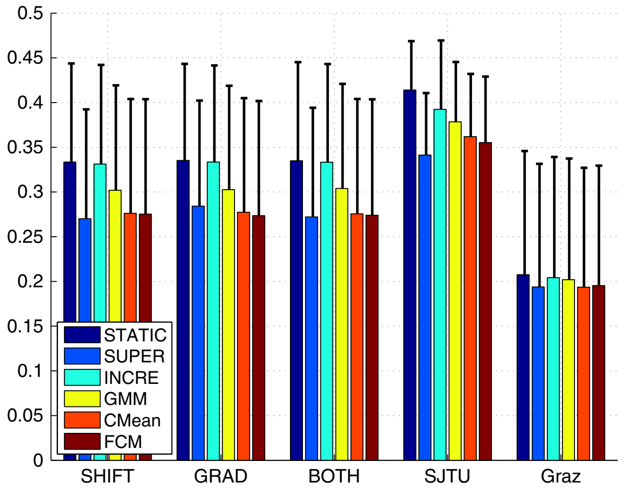


Figure 2. Error rates of six methods on different data types, including artificial data (SHIFT, GRAD, BOTH) and real data (SJTU, Graz). The results are averaged over all data sets with the same data type. Bars show the mean error rates, and lines indicate the standard deviations.

group, 10 data sets are constructed; therefore, we have 2250 data sets for each data type (SHIFT, GRAD, and BOTH).

In Figure 3, it can be seen how the separability and nonstationarity of the data affect the adaptation performance of different methods. For each fixed r_{cls} level, the error rates are averaged over all r_{chg} levels, and vice versa. When r_{chg} is big (or the nonstationary is severe), FCM shows a much lower error rate than INCRE and GMM methods, which means that FCM is more adaptive to changes than these two methods; but FCM does not outperform CMean when r_{chg} is very big. On the other hand, when r_{chg} is very small (or the data are quite stationary), none of the unsupervised methods show a noticeable improvement over STATIC. The reason may be that in this situation, a static classifier is good enough, whereas a complex unsupervised algorithm that needs parameter estimation will introduce computing error and result in a worse performance. From Figure 3(b), we can see that when r_{cls} is big (or the two classes are well separated), FCM achieves an error rate comparable to SUPER and CMean, which is much lower than INCRE and GMM. This means FCM can effectively discover the class information from unlabeled data, as long as the two classes are well separated.

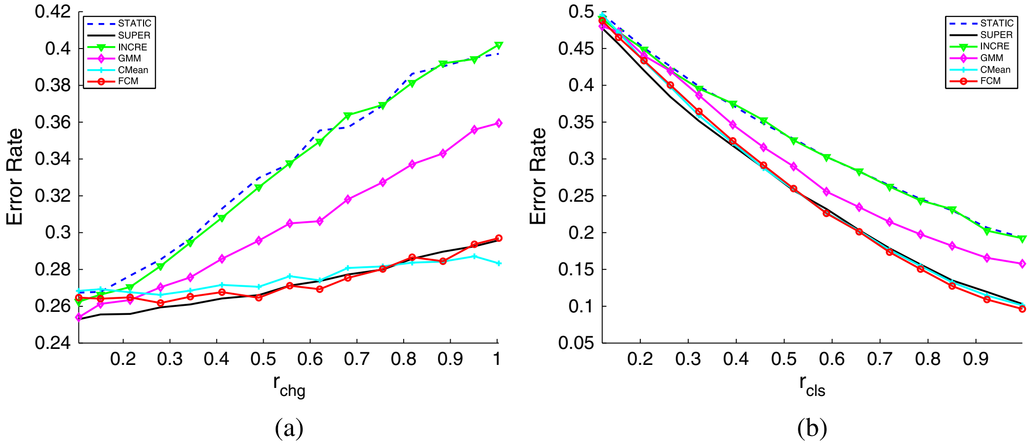


Figure 3. Effect of data properties on performances of six methods. The horizontal coordinate is r_{chg} or r_{cls}, and the vertical coordinate is error rate. Data type: BOTH. We do not show the plots for SHIFT and GRAD, as a similar performance can be found. (a) Effect of nonstationarity; (b) effect of separability.

DOI: 10.1002/acs

However, when r_{cls} is small (or the two classes are badly separated), all the unsupervised methods including FCM do not show evident improvement over STATIC. This means that if the data are essentially difficult to separate, then an advanced adaptation method will not help much. In such a case, more powerful feature extraction methods or even other BCI paradigms should be considered in order to achieve a satisfying result. In Figure 3, we only give the plot on data type BOTH, but not the plots on SHIFT and GRAD, as those plots show similar performances.

4.2. SJTU data and Graz data

Figure 4 shows the error rates of FCM versus the other five methods, when applied to real EEG data, including the SJTU data and the Graz data. Each point in the plot represents a data set, with the vertical axis meaning the error rate of FCM and the horizontal axis meaning the error rate of other methods. A point below the diagonal means a data set where FCM outperforms the other method (the error rate of FCM is lower). The SJTU data are plotted as crosses, and the Graz data are plotted as circles.

Paired t-test is used to check the statistical significance of differences between methods. The single-side p-value is shown in each subfigure of Figure 4. The p-value can be interpreted as the probability that the two methods have identical performances in nature, and a smaller p-value means the difference is more significant. Generally, $p = 0.05$ is the margin, and if $p < 0.05$, the difference can be deemed significant. From Figure 4, we can see the ranges of p-value for five sets of comparison are $[0, 0.0005]$, $[0, 0.025]$, $[0, 0.0005]$, $[0.01, 0.025]$, and $[0.1, 0.25]$, respectively. Except for Figure 4(e), all the p-values indicate the difference of comparison is significant, which shows a similar performance to that in Figure 2. This means the proposed method is effective for practical BCI applications. FCM outperforms STATIC, INCRE, and GMM obviously, but does not outperform SUPER. The advantage of FCM over CMean is not significant, which implies that the assumption of common mean change is nearly valid in most cases. As mentioned before, for the Graz data, the improvements of all the adaptive methods over STATIC is very small (within 1.5%). The reason may be that the Graz data are all quite stationary, with an r_{chg} less than 0.2 for most subjects. This also agrees with the performance shown in Figure 3(a).

In fact, statistical significance is often not very obvious in BCI algorithm comparison [37–39]. The main possible reason is that the sample size is not big enough. So, the mean and median classification accuracies are very different, which makes the use of some statistical methods such as ANOVA inappropriate [38]. Anyway, compared with previous research, the statistical significance presented in this work is acceptable.

Several hyper-parameters have to be determined before the application of the FCM algorithm, namely, the number of iterations N_{it}, the fuzziness exponent m, the number of historical estimations N_{hi}, and the number of unlabeled 'training' trials N_{ut}. The selection of these hyper-parameters and the effect of different selections on the averaged error rate of the SJTU data are shown in Figure 5. As mentioned before, N_{it} and m are predefined to be common for all the subjects, whereas N_{hi} and N_{ut} are subject specific. As Figure 5(a) shows, $N_{\text{it}} = 1$ is the best choice not only because it achieves the lowest error rate but also because it saves computational time. The degradation from $N_{\text{it}} = 1$ to $N_{\text{it}} = 2$ is significant. In fact, convergence of FCM has been proven by many ways in previous literature [40, 41]. FCM algorithm is convergent in itself with a limited number of trials, which coincides with our results, and we also find a big iteration number may cause overfitting and reduce the generality of the model. If the nonstationarity is severe and a sufficient number of trials are provided, a bigger N_{it} may be helpful, but the chance of divergence is increased if the data are not of unimodal distribution. This is an interesting research topic that may be further investigated. In our current work, limited number of trials is preferable because increased number cannot improve the performance. Coincidentally, less trials can guarantee the numerical convergence meanwhile.

On the other hand, the performance is not sensitive to the change of m within a certain range, as in Figure 5(b). Although $m = 1.6$ achieves the lowest error rate, we use $m = 2$ as the error rate is close to the lowest point and the computation is much simpler for $m = 2$. From Figure 5(c), we can see that although the best selection for N_{hi} is different from subject to subject, it seems $N_{\text{hi}} = 70 \sim 80$ is a good selection for all the data sets. $N_{\text{ut}} = 30$ is the best selection for about half of the data sets

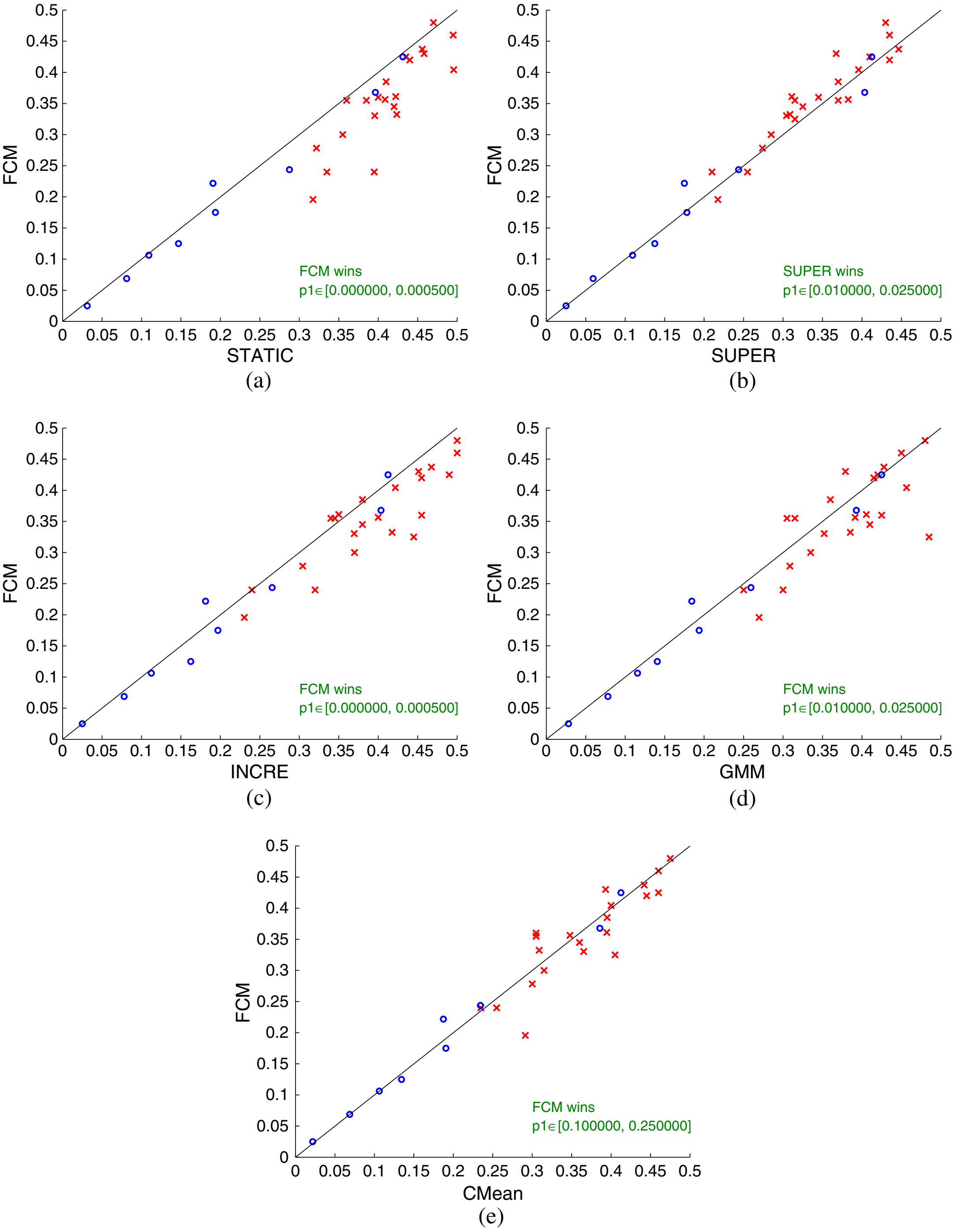
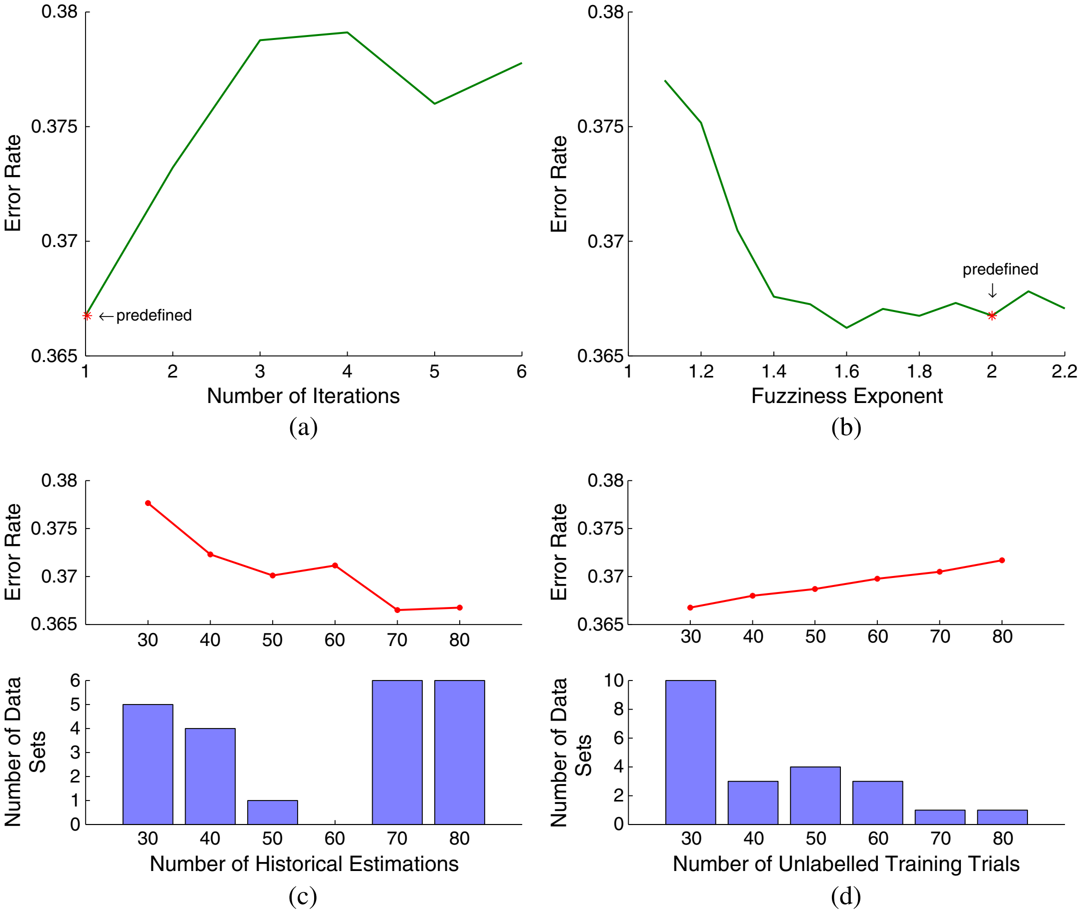


Figure 4. Error rates of FCM versus other methods for real data. (a) FCM versus STATIC; (b) FCM versus SUPER; (c) FCM versus INCRE; (d) FCM versus GMM; (e) FCM versus CMean. Each point in the plot represents a data set. A point below the diagonal means a data set where FCM outperforms the other method. Points plotted as circles are the Graz data and those in crosses are SJTU data.

(see Figure 5(d)), and it seems a bigger N_{ut} leads to a higher averaged error rate. This difference between N_{hi} and N_{ut} can be interpreted in such a way: the distribution of features changes over time; therefore, a smaller N_{ut} can sensitively reflect this change. On the other hand, the change of the estimations is much slower than the features; therefore, a bigger N_{hi} is appropriate.

DOI: 10.1002/acs



(a)

(b)

(c)

(d)

Figure 5. Selection of hyper-parameters in FCM and effects of hyper-parameters on averaged error rate of SJTU data. In each subfigure, one hyper-parameter is investigated, with other hyper-parameters fixed. (a) Effect of N_{it}; (b) effect of m; (c) selection and effect of N_{hi}; (d) selection and effect of N_{ut}.

From the parameter tuning, we can easily find such a contradiction: in order to sensitively reflect the change, time constant (such as N_{ut} in this paper) should be small; on the other hand, in order to achieve a credible estimation of the model parameters, time constant should be big. One possible approach to moderate this contradiction is to use a relatively simpler model with fewer model parameters.

For the Graz data, the CSP matrix is computed on a subject-specific time window, and multiple LDA classifiers are used at different time points to give a continuous output. During the simulated online adaptation, these classifiers are updated trial by trial independently. In other words, classifiers at different time points within a single trial are different and independent to each other. The similarity and difference among these classifiers can be considered as 'the adaptation within a single trial'. As the focus of this work targets the adaptation between trials, we do not investigate this problem in detail here.

5. CONCLUSION

The nonstationarity of EEG signals is an important issue in the research of BCI. Various supervised adaptation methods have been reported to overcome this problem. However, in practical application, the real intention of the subject is not always known to the system, and unsupervised adaptation methods are needed. So far, unsupervised adaptation methods for BCI have been reported by only a few articles. In this paper, we adopt the FCM algorithm for the unsupervised online (simulated)

DOI: 10.1002/acs

adaptation of a BCI application. We checked the performance of this method on both artificial data and real EEG data and investigated how the separability and nonstationarity of the data influence the adaptation performance. The proposed method achieves a better performance than other three existing unsupervised methods for most data types and shows a comparable performance to supervised adaptation. This confirms the effectiveness of our method. In addition, the results of real data agree with the analysis based on artificial data, which also confirms the validity of the artificially constructed data.

In an unsupervised scenario, a simple model with fewer parameters may have some advantages. In the GMM method, two class means and covariances have to be estimated in the iterative procedure, whereas in the FCM method, only two class means need to be iteratively estimated; the CMean model is even simpler, where only the common mean of the two classes is estimated. To a large extent, the success of FCM and CMean should be due to their simple models.

Another problem in adaptive BCIs is the adaptation of the feature extractor, for example, the W matrix in CSP. Unsupervised adaptation of the feature extractor is much more difficult than unsupervised adaptation of the classifier, as once the feature extractor changes, the new computed features will be very different from the old features. Inappropriate adaptation could cause confusion of the features. Therefore, the adaptation of the feature extractor should be more cautious. We will investigate this problem in the future.

ACKNOWLEDGEMENTS

This work is supported by the National Basic Research Program (973 Program) of China (no. 2011CB013305), National Natural Science Foundation of China (no. 51075265), and Science and Technology Commission of Shanghai Municipality (no. 09JC1408400).

REFERENCES

1. Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain–computer interfaces for communication and control. *Clinical Neurophysiology* 2002; **113**(6):767–791.
2. Millan JR. On the need for on-line learning in brain–computer interfaces. *Proceeding of International Joint Conference on Neural Networks*, Budapest, Hungary, July 2004. (IDIAP-RR03-30).
3. Kawanabe M, Krauledat M, Blankertz B. A Bayesian approach for adaptive BCI classification. *Proceedings of the 3rd International Brain–Computer Interface Workshop and Training Course*, Graz, Austria, 2006; 54–55.
4. Shenoy P, Krauledat M, Blankertz B, Rao RPN, Müller KR. Towards adaptive classification for BCI. *Journal of Neural Engineering* 2006; **3**(1):R13–R23.
5. Penny WD, Roberts SJ, Curran EA, Stokes MJ. EEG-based communication: a pattern recognition approach. *IEEE Transactions on Rehabilitation Engineering* 2000; **8**(2):214–215.
6. Tomioka R, Hill J, Blankertz B, Aihara K. Adapting spatial filtering methods for nonstationary BCIs. *Proceedings of Workshop on Information-Based Induction Sciences*, Osaka, Japan, 2006; 65–70.
7. Vidaurre C, Schlogl A, Cabeza R, Scherer R, Pfurtscheller G. A fully on-line adaptive BCI. *IEEE Transactions on Biomedical Engineering* 2006; **53**(6):1214–1219.
8. Sykacek P, Roberts SJ, Stokes M. Adaptive BCI based on variational Bayesian Kalman filtering: an empirical evaluation. *IEEE Transactions on Biomedical Engineering* 2004; **51**(5):719–727.
9. Sykacek P, Roberts S, Stokes M, Curran E, Gibbs M, Pickup L. Probabilistic methods in BCI research. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 2003; **11**(2):192–194.
10. Blankertz B, Kawanabe M, Tomioka R, Hohlefeld F, Nikulin V, Müller KR. Invariant common spatial patterns: alleviating nonstationarities in brain–computer interfacing. *Advances in Neural Information Processing Systems* 2008; **20**:113–120.
11. Vidaurre C, Schlogl A, Blankertz B, Kawanabe M, Müller KR. Unsupervised adaptation of the LDA classifier for brain–computer interfaces. In *Proceedings of the 4th International Brain–Computer Interface Workshop and Training Course*, Vol. 2008, 2008; 122–127.
12. Liu H, Wang J, Zheng C. Using self-organizing map for mental tasks classification in brain–computer interface. *Lecture notes in computer science* 2005; **3497**:327–332.
13. Buttfield A, Ferrez PW, del RMillan J. Towards a robust BCI: error potentials and online learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 2006; **14**(2):164–168.
14. Blankertz B, Curio G, Muller KR. Classifying single trial EEG: towards brain computer interfacing. *Advances in Neural Information Processing Systems* 2002; **1**:157–164.
15. Pfurtscheller G, Neuper C, Guger C, Harkam W, Ramoser H, Schlogl A, Obermaier B, Pregenzer M. Current trends in Graz brain–computer interface (BCI) research. *IEEE Transactions on Rehabilitation Engineering* 2000; **8**(2):216–219.

DOI: 10.1002/acs

16. Birbaumer N, Hinterberger T, Kubler A, Neumann N. The thought-translation device (TTD): neurobehavioral mechanisms and clinical outcome. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 2003; **11**(2):120–123.
17. Millan JR, Mouriño J. Asynchronous BCI and local neural classifiers: an overview of the adaptive brain interface project. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 2003; **11**(2):159–161.
18. Vidaurre C, Schlogl A, Cabeza R, Scherer R, Pfurtscheller G. Study of on-line adaptive discriminant analysis for EEG-based brain computer interfaces. *IEEE Transactions on Biomedical Engineering* 2007; **54**(3):550–556.
19. Sugiyama M, Krauledat M, Müller KR. Covariate shift adaptation by importance weighted cross validation. *The Journal of Machine Learning Research* 2007; **8**:985–1005.
20. Tsui CSL, Gan JQ. Comparison of three methods for adapting LDA classifiers with BCI application. *the 4th International Workshop on Brain–Computer Interfaces*, Graz, Austria, 2008; 116–121.
21. Lu S, Guan C, Zhang H. Unsupervised brain computer interface based on inter-subject information. *The 30th Annual International IEEE EMBS Conference*, Vancouver, Britsh Columbia, Canada, 2008; 638–641.
22. Gan JQ. Self-adapting BCI based on unsupervised learning. *Proceedings of the 3rd International Workshop on Brain–Computer Interfaces*, Graz, Austria, 2006; 50–51.
23. Eren SE, Grosse-Wentrup M, Buss M. Unsupervised classification for non-invasive brain–computer interfaces. *Proceedings of Automed Workshop*, VDI Verlag, Dusseldorf, Germany, 2007; 65–66.
24. Hasan B, Gan JQ. Unsupervised adaptive GMM for BCI. *Proceedings of the 4th International IEEE EMBS Conference on Neural Engineering*, Antalya, Turkey, 2009; 295–298.
25. Blumberg J, Rickert J, Waldert S, Schulze-Bonhage A, Aertsen A, Mehring C. Adaptive classification for brain computer interfaces. In *Proceedings of the 29th Annual International Conference of the IEEE EMBS, Lyon, France* 2007; **1**:2536–2539.
26. Liu G, Huang G, Meng J, Zhang D, Zhu X. Improved GMM with parameter initialization for unsupervised adaptation of brain–computer interface. *International Journal for Numerical Methods in Biomedical Engineering* 2010; **26**(6):681–691.
27. Mohamed NA, Ahmed MN, Farag A. Modified fuzzy C-mean in medical image segmentation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99)* 1999; **6**:3429–3432.
28. Kirlangic ME, Henning G. *EEG-Biofeedback and Epilepsy: Concept, Methodology and Tools for (Neuro) Therapy Planning and Objective Evaluation*. PhD Thesis, Ilmenau University of Technology, Germany, 2005.
29. Chen Y, Zhang L, Zhang D, Zhang D. Wrist pulse signal diagnosis using modified Gaussian models and fuzzy C-means classification. *Medical Engineering & Physics* 2009; **31**(10):1283–1289.
30. Liu G, Huang G, Meng J, Zhang D, Zhu X. Unsupervised adaptation based on fuzzy C-means for brain–computer interface. *The 1st International Conference on Information Science and Engineering (ICISE), Nanjing, China,* 2009:4122–4125.
31. Fukunaga K. *Introduction to Statistical Pattern Recognition*. Academic Press: San Diego, CA, 1990.
32. Müller-Gerking J, Pfurtscheller G, Flyvbjerg H. Designing optimal spatial filters for single-trial EEG classification in a movement task. *Clinical Neurophysiology* 1999; **110**(5):787–798.
33. Duda RO, Hart PE, Stork DG. *Pattern Classification*, (2nd edn). Wiley: New York, 2001.
34. Bezdek JC. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers Norwell: MA, USA, 1981.
35. Jasper HH. The ten-twenty electrode system of the international federation in electroencephalography and clinical neurophysiology. *Electroencephalography and Clinical Neurophysiology* 1958; **10**:371–375.
36. Leeb R, Lee F, Keinrath C, Scherer R, Bischof H, Pfurtscheller G. Brain–computer communication: motivation, aim, and impact of exploring a virtual apartment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 2007; **15**(4):473–482.
37. Lotte F, Guan C. Spatially regularized common spatial patterns for EEG classification. *2010 IEEE International Conference on Pattern Recognition*, 2010; 3712–3715.
38. Lotte F, Guan C. Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms. *IEEE Transactions on Biomedical Engineering* 2011; **58**(2):355–362.
39. Arvaneh M, Guan C, Ang K, Quek H. Optimizing the channel selection and classification accuracy in EEG-based BCI. *IEEE Transactions on Biomedical Engineering* 2011; **58**(6):1865–1873.
40. Hathaway RJ, Bezdek JC. Recent convergence results for the fuzzy C-means clustering algorithms. *Journal of Classification* 1988; **5**(2):237–247.
41. Groll L, Jakel J. A new convergence proof of fuzzy C-means. *IEEE Transactions on Fuzzy Systems* 2005; **13**(5):717–720.

DOI: 10.1002/acs