INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN BIOMEDICAL ENGINEERING
Int. J. Numer. Meth. Biomed. Engng. 2010; **26**:681–691
Published online 1 December 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cnm.1362

Improved GMM with parameter initialization for unsupervised adaptation of Brain–Computer interface

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

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

SUMMARY

An important property of brain signals is their nonstationarity. How to adapt a brain–computer interface (BCI) to the changing brain states is one of the challenges faced by BCI researchers, especially in real application where the subject's real intent is unknown to the system. Gaussian mixture model (GMM) has been used for the unsupervised adaptation of the classifier in BCI. In this paper, a method of initializing the model parameters is proposed for expectation maximization-based GMM parameter estimation. This improved GMM method and other two existing unsupervised adaptation methods are applied to groups of constructed artificial data with different data properties. Performances of these methods in different situations are analyzed. Compared with the other two unsupervised adaptation methods, this method shows a better ability of adapting to changes and discovering class information from unlabelled data. The methods are also applied to real EEG data recorded in 19 experiments. For real data, the proposed method achieves an error rate significantly lower than the other two unsupervised methods. Results of the real data agree with the analysis based on the artificial data, which confirms not only the effectiveness of our method but also the validity of the constructed data. Copyright © 2009 John Wiley & Sons, Ltd.

Received 20 May 2009; Revised 14 October 2009; Accepted 20 October 2009

KEY WORDS: brain–computer interface (BCI); electroencephalogram (EEG); unsupervised adaptation; Gaussian mixture model (GMM); expectation maximization (EM)

1. INTRODUCTION

People with severe neuromuscular disorders or suffering from a locked-in syndrome need alternative methods for communication and control. Brain–computer interfaces (BCIs) are systems that allow their users to communicate with a computer program or control a mechanical device directly by intent rather than by neuromuscular pathway [1]. One of the challenges faced by BCI researchers is the nonstationarity of the subject's brain states, reflected as changes in the statistical properties of the electroencephalogram (EEG) signals, which has been reported in the literature [2–5]. The nonstationarity may be caused by various reasons, e.g. changes in the concentration or excitation

*Correspondence to: Xiangyang Zhu, State Key Laboratory of Mechanical System and Vibration, Shanghai Jiao Tong University, Shanghai, China.
†E-mail: mexyzhu@sjtu.edu.cn

Contract/grant sponsor: National Natural Science Foundation; contract/grant number: 50525517
Contract/grant sponsor: Science and Technology Commission of Shanghai Municipality; contract/grant numbers: 08JC1412100, 09JC1408400

level, changes in the mental task involved, influence of feedback, demands for visual processing, fatigue, artifacts such as swallowing and blinking, changes in the impedance or even position of the electrodes, and ambient noise [2–4, 6–13]. Owing to the nonstationarity, the BCI system trained on one session may become ill-suited to the succeeding sessions.

In order to improve the adaptability of a BCI system to the changing brain states, various adaptation methods have been investigated. In [14–16] the classifier was updated manually after several runs, according to the experimenter's experience and judgement. In [2–4, 13, 17–19] 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.

However, as pointed out in [2, 6, 11], in a practical application scenario, the real intention of the subject is not always known to the system. Considering this situation, some research groups have paid attention to unsupervised adaptation algorithms. In [11] three types of unsupervised adaptation procedures for linear discriminant analysis (LDA) classifier were proposed and analyzed. A limit of these procedures is that they are based on an assumption that during the online sessions, the vector connecting the two class means keeps nearly constant, which may not always be true in real BCI. Without such assumptions, Gan [20] 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.

To avoid such disadvantages, some non-incremental unsupervised clustering methods have also been reported. In [21], Eren used Gaussian mixture model (GMM) for unsupervised clustering of EEG patterns, but this work did not check the online adaptation performance of the method. In [22], Blumberg proposed an adaptive linear discriminant analysis (ALDA) method for simulated online clustering of EEG patterns, in which expectation maximization (EM) method was used to estimate the means and a common covariance of the two classes. This method is actually a GMM-based adaptation method, although not explicitly named. The initial data distributions were estimated from a labelled training period. A limit of this strategy 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.

In this paper, the model parameters in GMM are estimated by the EM algorithm combined with an initialization method. This improved GMM (iGMM) method is used to update the LDA classifier for a simulated online BCI scenario. In order to investigate its performance in different situations, the proposed method is applied to several types of constructed artificial data with different properties, as well as real EEG data collected from 19 experiments. The performance of our method is compared with a static LDA which is not updated, a supervised adaptive LDA, an incrementally updated LDA, and a GMM-based adaptive LDA. Our method shows a better performance than the other two unsupervised adaptation methods not only on the artificial data, but also on the real EEG data. We also investigate how the data properties (separability and nonstationarity) affect the performance of different adaptation methods. Analysis on artificial data shows that our method is more adaptive to nonstationarity, and can more effectively discover the class information from unlabelled data. Results of the real data show similar phenomenon to the artificial data. Compared with other unsupervised methods, our method has some advantages. First, the classifier is not updated incrementally, but recomputed from the recent unlabelled trials every time, which makes the method robust to bad trials. In other words, if one trial is incorrectly classified, it will not affect the classifier significantly, since the classifier is updated based on a number of recent trials. Second, the initial parameter set (mean and covariance of the two classes) used in the iGMM is not constant, but updated along time according to the historical estimations, which makes it adaptive to significant changes in the data distribution during long-term use.

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

DOI: 10.1002/cnm

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 variance for one class and at the same time minimize variance for the other class. Mathematically, CSP is realized by simultaneous diagonalization of the covariance matrices for the two classes [23]. The formulas for this algorithm can be seen in [24]. The normalized log-variances of six most discriminative components were used as features. The transformation to logarithmic values makes the distribution of the features approximately normal.

2.2. LDA classifier

Fisher's LDA [25] 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

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

and the bias is

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

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

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

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

2.3. Supervised adaptation of the classifier

We denote the parameter set as $\Theta=\{\mu_1,\mu_2,\Sigma_1,\Sigma_2\}$. Since 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 $x(t)$ is the latest feature vector whose true label is already known, we update Θ in the following way:

$$\mu_i(t)=\mu_i(t-1)\cdot(1-UC)+x(t)\cdot UC \tag{4}$$

$$\Sigma_i(t)=\Sigma_i(t-1)\cdot(1-UC)+(x(t)-\mu_i(t))\cdot(x(t)-\mu_i(t))^{\mathrm{T}}\cdot UC \tag{5}$$

where i is the true label of $x(t)$, and UC is the update coefficient.

2.4. Unsupervised adaptation of the 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. In this paper, we compare our iGMM-based method with two existing unsupervised adaptation methods, namely, an incremental adaptation method, and a GMM-based method. The adaptation of the feature extractor is another important topic, which is not considered in this work.

2.4.1. Incremental adaptation. In [20], an incremental updating method is proposed to update the mean and covariance of each class. This method consists of two steps:

Step 1: when a new feature vector x comes, its label is estimated by unsupervised clustering;

Step 2: the mean and covariance of the class into which the new x is classified are updated in a way similar to equations (4) and (5).

DOI: 10.1002/cnm

It can be seen that, this method is a 'decision-directed' approach [25], 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 [20], this problem is described as 'when to adapt', and Gan mentioned that some safety precautions should be taken, e.g. to check the confidence level, or to check whether there exists an error potential. In this paper, we use the following way to check the confidence level of the classification result: if the absolute value of the LDA classifier output $y = w^{\mathrm{T}} \cdot x + b$ exceeds a threshold y_{th}, we say the classification is confident enough, and update the classifier (by updating the parameter set Θ) using this new feature; otherwise make no change.

In this method, Θ_{new} is got by modifying Θ_{old} with an incremental value. We name this method as incremental adaptation method, in contrast with the methods below.

2.4.2. GMM-based adaptation. Based on the assumption that the feature vectors of the two classes are normally distributed in feature space, Eren proposed a GMM-based method for unsupervised clustering of EEG patterns in [21], in which the parameter set Θ of the model was estimated by the well-known EM algorithm. But this work did not check the online adaptation performance of the proposed method. In [22], Blumberg proposed an ALDA method, which is actually a GMM-based adaptation method, to update the LDA classifier in a simulated online manner. In this method, Θ_{new} at every step is estimated by the EM algorithm on the basis of an initial value Θ_{init} and the recent unlabelled trials, but not on Θ_{old}.

The EM algorithm is an iterative algorithm for parameter estimation. Every iteration of EM method consists of two steps, namely E-step and M-step.

E-step: Compute the expected classes for each feature vector:

$$\begin{aligned} P(\omega_i | x_k, \Theta_t) &= \frac{p(x_k | \omega_i, \Theta_t) \cdot P(\omega_i | \Theta_t)}{p(x_k | \Theta_t)} \\ &= \frac{p(x_k | \omega_i, \mu_i(t), \Sigma_i(t)) \cdot P(\omega_i)}{\sum_{j=1}^{2} p(x_k | \omega_j, \mu_j(t), \Sigma_j(t)) \cdot P(\omega_j)} \end{aligned} \tag{6}$$

where x_k means the kth feature vector, $\Theta_t = \{\mu_1(t), \mu_2(t), \Sigma_1(t), \Sigma_2(t)\}$ means the estimated Θ on the tth iteration, and $i \in \{1, 2\}$ means class label.

M-step: Given the data's class membership distributions, compute new Θ that maximizes the likelihood:

$$\begin{aligned} \mu_i(t+1) &= \frac{\sum_k P(\omega_i | x_k, \Theta_t) \cdot x_k}{\sum_k P(\omega_i | x_k, \Theta_t)} \\ \Sigma_i(t+1) &= \frac{\sum_k P(\omega_i | x_k, \Theta_t) \cdot C_{i,k}}{\sum_k P(\omega_i | x_k, \Theta_t)} \end{aligned} \tag{7}$$

where $C_{i,k} = (x_k - \mu_i(t+1)) \cdot (x_k - \mu_i(t+1))^T$, and k is the index of trial. If I_{cu} is the current trial index, and N_{ut} is the size of the unlabelled 'training' set used to update the classifier, then $k = I_{cu} - N_{ut} + 1, \ldots, I_{cu}$. Since EM algorithm is iterative, we need to determine the number of iterations T_{it}. In our computational experiment, we let $T_{it} = 1$ since we find one iteration is sufficient for the adaptation, and more iterations cannot further improve the performance.

2.4.3. iGMM-based adaptation. A crucial issue in iterative parameter estimation is to determine the initial parameter value Θ_{init} for the first iteration. In [22] the Θ_{init} is computed from the labelled training data set. A limit of this strategy is that, if the feature distribution shifts far away from the training session during long-term use, this Θ_{init} will not be valid any more. In order to solve this problem and improve the adaptability of the classifier, we propose the iGMM based approach.

In this iGMM-based approach, we treat Θ_{init} as a variational parameter but not constant. The current Θ_{init} is estimated based on the historical estimations of Θ. Let $\Theta^{(0)}$ denote the Θ computed from the labelled training session, $\Theta_{\mathrm{init}}^{(k)}$ denote the Θ_{init} used for estimation when the kth unlabelled

DOI: 10.1002/cnm

trial come, and $\Theta^{(k)}$ denote the estimated Θ after the kth unlabelled trial. Then we determine $\Theta_{\text{init}}^{(k)}$ in the following way:

$$\Theta_{\text{init}}^{(k)} = \begin{cases} \Theta^{(0)} & \text{if } k \leqslant N_{\text{hi}} \\ \dfrac{1}{N_{\text{hi}}} \displaystyle\sum_{j=k-N_{\text{hi}}}^{k-1} \Theta^{(j)} & \text{else} \end{cases} \tag{8}$$

where N_{hi} is the size of the historical set of the estimations. Once Θ_{init} is determined, the succeeding calculations are the same as in GMM-based approach.

3. APPLICATION

3.1. Artificial data

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

As reported in [4], there are mainly two types of change 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 change (BOTH).

Since the main focus of this work is the adaptation of the classifier, we directly construct the feature vectors rather than EEG signals. The artificial features are constructed in such a procedure: (1) Two classes of static feature vectors are constructed, with each class following a multi-dimensional Gaussian distribution. The distance between the two class centers is controlled to determine the separability of the data. (2) The features are divided into three sessions, with the first one as labelled training session and the last two as unlabelled test sessions. (3) A bias is added to the test 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.

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

$$r_{\text{cls}} = \frac{|\mu_1 - \mu_2|\sqrt{n_1 n_2}}{\sqrt{\left|(\Sigma_1 + \Sigma_2) \cdot \dfrac{(\mu_1 - \mu_2)}{|\mu_1 - \mu_2|}\right|}(n_1 + n_2)} \tag{9}$$

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

$$r_{\text{chg}} = \frac{|\mu_{\text{tr}} - \mu_{\text{te}}|\sqrt{n_{\text{tr}} n_{\text{te}}}}{\sqrt{\left|(\Sigma_{\text{tr}} + \Sigma_{\text{te}}) \cdot \dfrac{(\mu_{\text{tr}} - \mu_{\text{te}})}{|\mu_{\text{tr}} - \mu_{\text{te}}|}\right|}(n_{\text{tr}} + n_{\text{te}})} \tag{10}$$

where the subscripts tr and te indicate the labelled training data set and the unlabelled test data set, respectively. This parameter can be interpreted as the difference between the training data and

DOI: 10.1002/cnm

the test data. A bigger r_{chg} means that the test data are more different from the training data, i.e. 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.

3.2. Experimental setup and real data

3.2.1. Subjects. Seven right-handed subjects (five male and two female, age 23–27 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 or right hand movement. The arrowed cross was shown until the end of second 5. During this time 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 two kinds of movements were decided by the subject herself/himself, e.g. 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, i.e. 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 flat. 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. This was to alleviate fatigue and to avoid the subject getting exhausted. Data in one data set were recorded on the same day. Each subject took one to three experiments, resulting in a total of 19 data sets.

During the online experiments, the classifier actually used to give a feedback was an LDA classifier that 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 done 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.

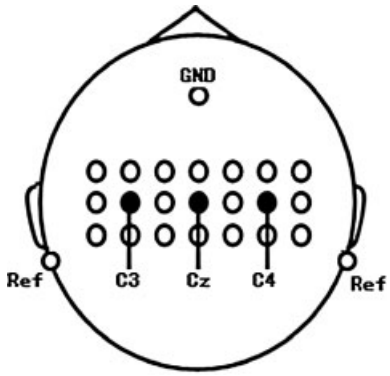

Figure 1. Placement of the 21 EEG electrodes. The three electrodes painted black are C3, Cz, and C4, respectively. 'GND' means ground electrode and 'Ref' means reference electrodes.

DOI: 10.1002/cnm

3.2.3. Recordings. EEG signals were recorded using a SynAmps system (Neuroscan, U.S.A.). 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 [26, 27] (see Figure 1). Horizontal and vertical 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 5–30 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.

4. RESULTS AND DISCUSSION

For both artificial and real data, we check the performances of five methods, namely, a static LDA classifier without adaptation (STAT), supervised adaptation (SUPER), incremental adaptation (INCRE), GMM-based adaptation (GMM), and the proposed iGMM-based approach (iGMM). Hyper-parameters in each methods are optimized based on the training data.

The artificial data have the same framework as the real data, i.e. three sessions for each data set, with the first one as training session and the other two as testing sessions. During the hyper-parameter tuning phase, the training session is further divided into two even parts: training part and validation part. During the simulated online application, every trial in the test sessions is provided with an estimated label by each method. The error rate of each method is calculated as the ratio of incorrectly classified trials to the total test trials.

4.1. Artificial data

By varying r_{cls} and r_{chg} each among 11 degrees, we get 121 groups of data sets, each group with fixed r_{cls} and r_{chg} level. In each group 10 data sets are constructed; therefore, we have 1210 data sets for each data type (SHIFT, GRAD, and BOTH).

Figure 2 shows the averaged error rates of all the methods on the three types of artificial data. We also present the result for real data (indicated as REAL) for the purpose of comparison. The error rates in this figure are averaged over all data sets with the same data type. From this figure we

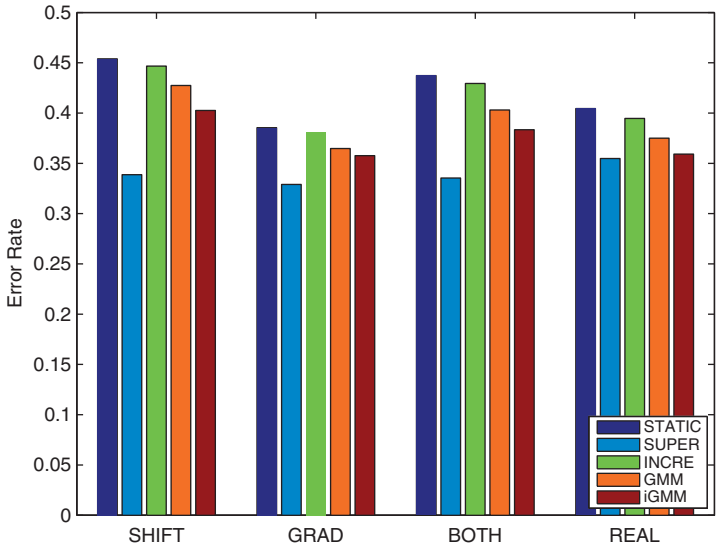

Figure 2. Error rates of the methods on different types of artificial data, as well as on the real data. The results are averaged over all data sets with the same data type.

DOI: 10.1002/cnm

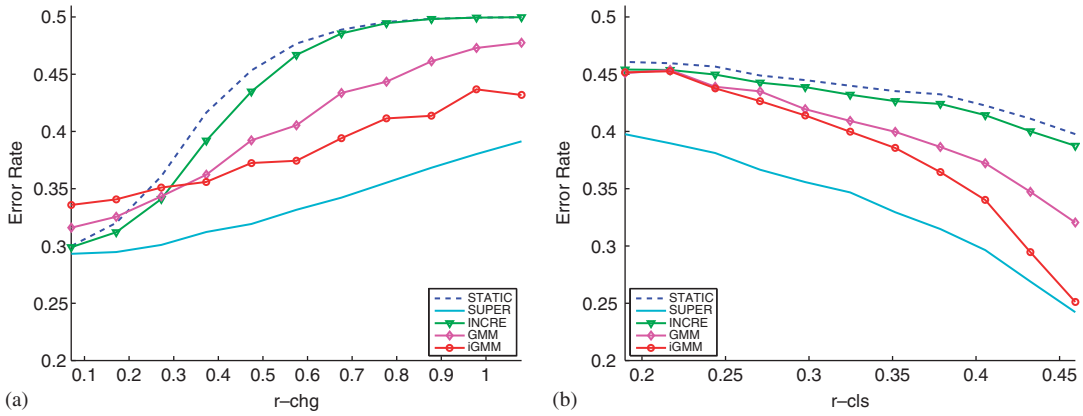

Figure 3. Effect of data properties on the performance of each method. The horizontal axis is r_{chg} or r_{cls}, and the vertical axis is error rate. Data type BOTH. We do not show the plots for SHIFT and GRAD, since a similar phenomenon can be found. (a) Effect of nonstationarity and (b) effect of separability.

can see that, for all data types including REAL, our method shows a significant better performance than STATIC and INCRE, a considerable improvement over GMM. However, iGMM does not achieve a perfect performance comparable to SUPER. One possible reason that impedes iGMM from further improving may be that there are too many parameters (class means and covariances) to estimate with a limited number of trials. Simpler models with fewer model parameters may have some advantages, such as the Fuzzy C-Means (FCM) method in which only the class means and a predefined fuzziness exponent m are used to describe the model, but no covariance is used.

Figure 3 shows how the separability and nonstationarity of the data affect the adaptation performance of different methods. For each fixed r_{cls} level, the error rates were averaged over all r_{chg} levels, and vice versa. From Figure 3(a) we can see that, when r_{chg} is big (or the nonstationary is severe), iGMM shows a much lower error rate than other unsupervised methods, which means that iGMM is more adaptive to changes. However, when r_{chg} is very small (or the data are quite stationary), all the unsupervised methods show a higher error rate than STATIC, especially iGMM. This is because in this situation, a static classifier is good enough, while 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 enough (or the two classes are well separated), iGMM achieves an error rate close to SUPER, which is much lower than other unsupervised methods. This means that our method can effectively discover the class information from unlabelled data, as long as the two classes are well separated. However, when r_{cls} is small (or the two classes are badly separated), all the unsupervised methods including our iGMM do not show evident improvement over STATIC. This means if the data are essentially difficult to separate, then an advanced adaptation method will not help a lot. In such a case, more powerful feature extraction methods or even other BCI paradigms should be considered in order to achieve a satisfying result. Figure 3 only gives the plot of data type BOTH, but not the plots of SHIFT and GRAD, since those plots show a similar phenomenon.

4.2. Real data

Figure 4 shows the error rates of iGMM versus the other four methods, when applied to real EEG data. Each point in the plot represents a data set, with the vertical axis meaning the error rate of iGMM, and the horizontal axis meaning the error rate of other methods. A point below the diagonal means a data set where iGMM outperforms the other method. Points plotted as circles are the four data sets in Table I, which will be discussed later. From Figure 4 we can see a similar phenomenon as in Figure 2. This means our method is effective for practical BCI applications.

DOI: 10.1002/cnm

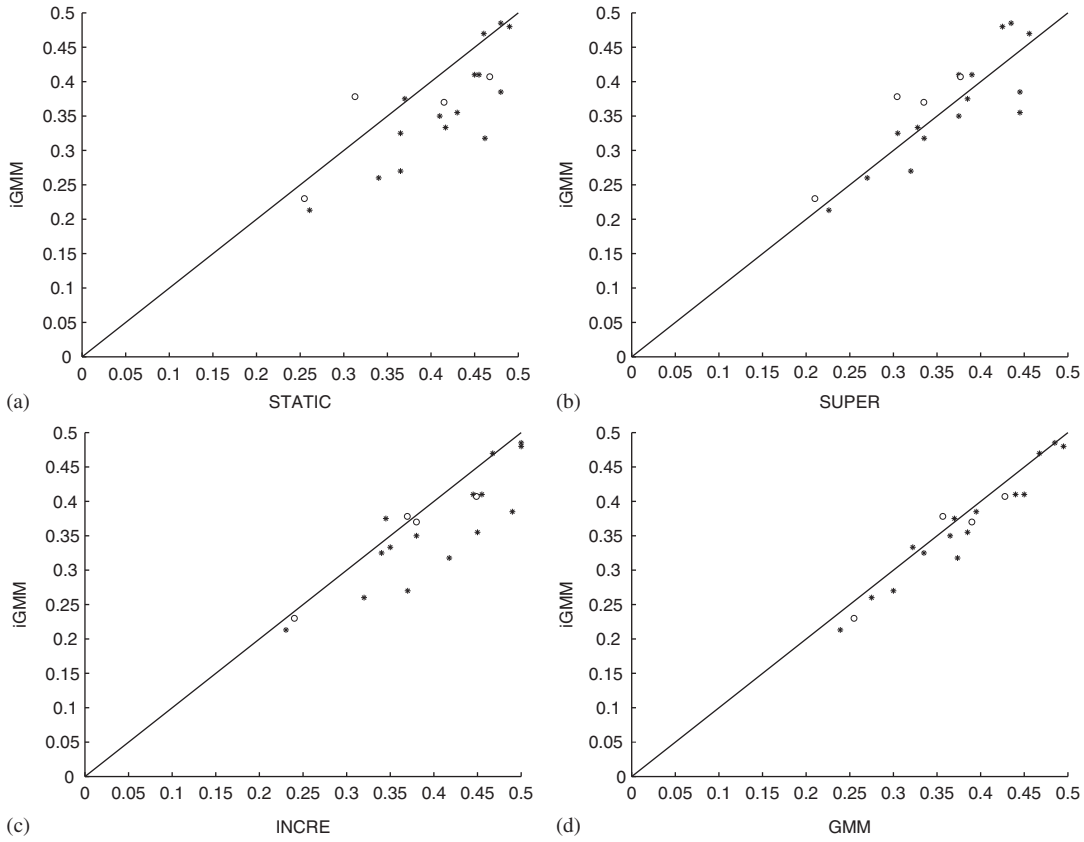

Figure 4. Error rates of iGMM versus other methods for real data. (a) iGMM versus STATIC; (b) iGMM versus SUPER; (c) iGMM versus INCRE; and (d) iGMM versus GMM. Each point in the plot represents a data set. A point below the diagonal means a data set where iGMM outperforms the other method. Points plotted as circles are the four data sets in Table I.

Table I. The data properties and the error rates of several typical real data sets, which are plotted as circles in Figure 4.

Data	r_{cls}	r_{chg}	STAT	SUPER	INCRE	GMM	iGMM
1	0.66	0.52	0.255	0.210	0.240	0.255	**0.230**
2	0.38	0.11	0.313	0.304	0.370	**0.356**	0.378
3	0.20	0.64	0.467	0.377	0.449	0.428	**0.407**
4	0.22	0.18	0.415	0.335	0.380	0.390	**0.370**

Bold font indicates the lowest error rate achieved by the three unsupervised methods.

In Table I, the error rates of four typical real data sets with different property configures are shown. Here we can check whether the relationship between data properties and error rates shown in Figure 3 still holds for real data or not. Data set 1 has a big r_{cls} and a big r_{chg}, which means the two classes are well separated and at the same time the nonstationarity is severe. As we expected, the error rates are low for all methods, and iGMM shows a better performance than other methods except SUPER. Data set 2 has a very small r_{chg}, which means the data are quite stationary. As we can see, all the three unsupervised methods show worse performance than STATIC, especially iGMM. Data set 3 has a small r_{cls} and a big r_{chg}, which means that the data are badly separated and are very nonstationary. Therefore, the whole performances are relatively poor, but iGMM still shows a better performance than the other unsupervised methods. Data set 4 has a small r_{cls} and a relatively small r_{chg}. As expected, the whole performances are poor. But surprisingly, iGMM still outperforms other unsupervised methods, may be because the r_{chg} is not so small as in data

DOI: 10.1002/cnm

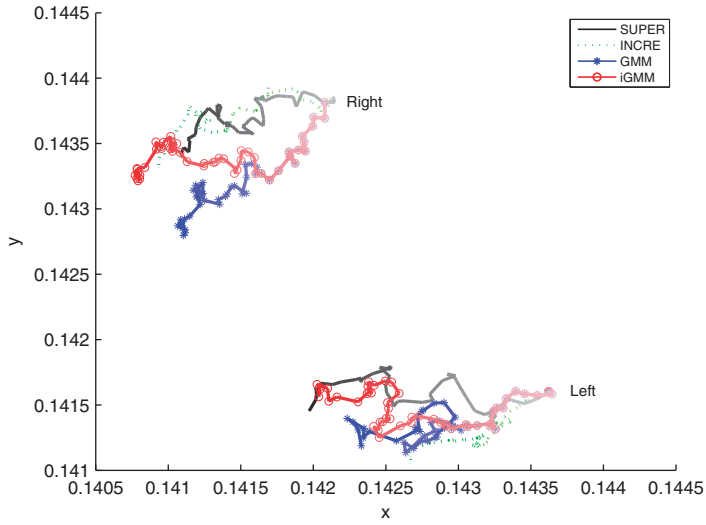

Figure 5. Time course of the class means estimated by different methods, projected on a 2-dimensional feature subspace. Time is indicated by darkness (from light to dark). Starting points of the trajectories are the class means used by STATIC. The data set is data 1 in Table I.

set 2. As a whole, these data sets agree with the phenomenon shown in Figure 3, which not only confirms the effectiveness of our proposed method, but also indirectly confirms the validity of the constructed artificial data.

In order to visually show how each method adapts to the change of the features, time course of the class means estimated by different methods is shown in Figure 5, in which time is indicated by darkness (from light to dark). The plot is got from data 1 in Table I. Since the exact distribution of the features is unavailable, we consider the estimation by SUPER as the best approximation. From Figure 5 it can be seen that, at the late stage of the time course, iGMM gives a closer estimation to SUPER than GMM does. At the early stage, iGMM and GMM give superposed trajectories, because when $k \leqslant N_{\text{hi}}$ (see Equation (8)), the Θ_{init} used in iGMM is the same as in GMM. It also should be noted that, although the trajectory of class right estimated by INCRE is close to that by SUPER, the trajectory of class left is poorly estimated by INCRE, because many left trials are misclassified into the right class.

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 research groups. In this paper, we proposed an iGMM based unsupervised adaptation approach for online BCI application, in which the GMM parameters were estimated by the EM algorithm combined with an initialization method. We checked the performance of this method on both constructed 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 two existing unsupervised methods on both the artificial data and real data. This confirms the effectiveness of our method. In addition, results of the real data agree with the analysis based on the artificial data, which also confirms the validity of the constructed artificial data.

As we pointed out previously, in this paper, we only studied the unsupervised adaptation of the classifier, but not the feature extractor. Our future work will focus on the unsupervised adaptation of the feature extractor and finally the whole BCI system.

DOI: 10.1002/cnm

ACKNOWLEDGEMENTS

This work was jointly supported by National Natural Science Foundation (Grant No. 50525517), and Science and Technology Commission of Shanghai Municipality (Grant No. 08JC1412100 and 09JC1408400).

REFERENCES

1. Wolpaw J, Birbaumer N, McFarland D, Pfurtscheller G, Vaughan T. Brain–computer interfaces for communication and control. *Clinical Neurophysiology* 2002; **113**(6):767–791.
2. Millan J. On the need for on-line learning in brain–computer interfaces. *Proceedings of the 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 R, Müller K. Towards adaptive classification for BCI. *Journal of Neural Engineering* 2006; **3**(1):13–23.
5. Penny W, Roberts S, Curran E, Stokes M. 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, November 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 S, 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 K. 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 K. Unsupervised adaptation of the LDA classifier for brain–computer interfaces. *Proceedings of the 4th International Brain–Computer Interface Workshop and Training Course* 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, vol. 3497. Springer: Berlin, 2005; 327–332.
13. Buttfield A, Ferrez P, del RMillan J. Towards a robust BCI: error potentials and online learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 2006; **14**(2):164.
14. Blankertz B, Curio G, Muller K. 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.
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 J, 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 K. Covariate shift adaptation by importance weighted cross validation. *The Journal of Machine Learning Research* 2007; **8**:985–1005.
20. Gan J. Self-adapting BCI Based on Unsupervised Learning. *Third International Workshop on Brain–Computer Interfaces*, Graz, Austria, 2006; 50–51.
21. Eren S, Grosse-Wentrup M, Buss M. Unsupervised classification for non-invasive brain–computer interfaces. *Proceedings of Automed Workshop*. VDI Verlag: Düsseldorf, Germany, October 2007; 65–66.
22. Blumberg J, Rickert J, Waldert S, Schulze-Bonhage A, Aertsen A, Mehring C. Adaptive classification for brain computer interfaces. *Conference Proceedings—IEEE Engineering in Medicine and Biology Society*. IEEE: New York, 2007; 2536–2539.
23. Fukunaga K. *Introduction to Statistical Pattern Recognition*. Academic Press: New York, 1990.
24. 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.
25. Duda R, Hart P, Stork D. *Pattern Classification* (2nd edn). Wiley: New York, 2001.
26. Jasper H. The ten-twenty electrode system of the international federation in electroencephalography and clinical neurophysiology. *EEG Journal* 1958; **10**:371–375.
27. Oostenveld R, Praamstra P. The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology* 2001; **112**(4):713–719.

DOI: 10.1002/cnm