INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF NEURAL ENGINEERING
J. Neural Eng. **3** (2006) R13–R23 doi:10.1088/1741-2560/3/1/R02

TUTORIAL

Towards adaptive classification for BCI*

Pradeep Shenoy[1,2], Matthias Krauledat[2,3], Benjamin Blankertz[2], Rajesh P N Rao[1] and Klaus-Robert Müller[2,3]

[1] Computer Science Department, University of Washington, Box 352350, Seattle, WA 98195, USA
[2] Fraunhofer FIRST (IDA), Kekuléstr. 7, 12 489 Berlin, Germany
[3] Department of CS, University of Potsdam, August-Bebel-Str. 89, 14 482 Potsdam, Germany

Received 19 October 2005
Accepted for publication 27 January 2006
Published 1 March 2006
Online at stacks.iop.org/JNE/3/R13

Abstract
Non-stationarities are ubiquitous in EEG signals. They are especially apparent in the use of EEG-based brain–computer interfaces (BCIs): (a) in the differences between the initial calibration measurement and the online operation of a BCI, or (b) caused by changes in the subject's brain processes during an experiment (e.g. due to fatigue, change of task involvement, etc). In this paper, we quantify for the first time such systematic evidence of statistical differences in data recorded during offline and online sessions. Furthermore, we propose novel techniques of investigating and visualizing data distributions, which are particularly useful for the analysis of (non-)stationarities. Our study shows that the brain signals used for control can change substantially from the offline calibration sessions to online control, and also within a single session. In addition to this general characterization of the signals, we propose several adaptive classification schemes and study their performance on data recorded during online experiments. An encouraging result of our study is that surprisingly simple adaptive methods in combination with an offline feature selection scheme can significantly increase BCI performance.

1. Introduction

The goal of a brain–computer interface (BCI) is to translate the intent of a subject directly into control commands for a computer application or a neuroprosthesis. This intent is estimated from brain signals measured via signals from the scalp or from invasive techniques, cf [6, 17, 32] for an overview. A significant challenge in designing a BCI is to balance the technological complexity of interpreting the user's brain signals with the amount of user training required for successful operation of the interface.

The BCI scenario involves two (possibly) adaptive parts, the user and the system. The operant conditioning approach [1, 11, 28] uses a fixed translation algorithm to generate a feedback signal from EEG. Users are not equipped with a mental strategy they should use. Rather, they are instructed to watch a feedback signal and to find out how to voluntarily control it. Successful operation is reinforced by a reward stimulus. In such BCI systems, the adaptation of the user is crucial and typically requires extensive training. On the other hand, machine learning techniques allow us to fit many parameters of a general translation algorithm to the specific characteristics of the user's brain signals [4, 22, 24, 25]. This is done by a statistical analysis of a calibration measurement in which the subject performs well-defined mental acts such as imagined movements [19, 27]. Here in principle no adaptation of the user is required, but it can be expected that users will adapt their behavior during feedback operation. The idea of the machine learning approach is that a flexible adaptation of the system relieves a good amount of the learning load from the subject. Most BCI systems are somewhere between those extremes. Every system reacts differently to the changes of brain activity of an adapting user. Here we examine the influence of non-stationary brain signals in the operation of the Berlin BCI (BBCI). This system represents the machine learning approach to BCI and has had considerable success in allowing user operation with bitrates of up to 35 bits per

* Part of the 3rd Neuro-IT and Neuroengineering Summer School Tutorial Series.

minute (bpm) and as little as 30 min of calibration/training, cf [3].

The central thesis of our BBCI design is that our offline classification accuracy, coupled with the reliability of the signals generated during motor imagery, should yield a classifier with accurate online performance and no learning on the part of the user. While the system indeed achieves good accuracies in online sessions with no user training, we observed in several cases that the performance can be enhanced by manually adjusting some parameters of the translation algorithm, such as bias or scaling of the classifier output. Further, during online sessions, subjects report phases where the accuracy of BCI control is degraded. Thus, there is considerable evidence of non-stationarity in the BCI classification problem.

Various approaches have been suggested for coping with this non-stationary behavior of EEG signals. In the BCI context, the large variety of methods used for control naturally lead to different schemes for adapting algorithm parameters during a BCI session. As a result, the success and applicability of the adaptation scheme used is heavily dependent on the chosen BCI scenario.

In [33], a visual BCI feedback was described in which the user was able to control a computer cursor in two dimensions, trying to hit one of the eight possible targets. The classification algorithm used two distinct band-power features acquired from a small subset of 64 scalp electrodes. Several scaling factors were used to translate these features into positions on the screen, four of which were successively adapted to the individual user during the session. Similarly, [20] investigated a scenario involving a four-class BCI classification problem. The estimation of means and covariance matrices for each of the classes was iteratively updated in a simulated online scenario; these parameter changes indicated the possibility of considerable improvement for online control. In this case, several channels from centroparietal scalp regions were used for the extraction of spectral features. In another offline study, this finding was backed by [31]; here, the parameters of a quadratic classifier (QDA) were adapted after each trial of a cursor-movement task. After a careful update parameter selection, the resulting classification was superior to the static classifier that was used from the start.

In each of these studies, the used method of adaptivity differs slightly and it is hard to transfer these results to other classification approaches, since the underlying changes in the models might differ. Also, this body of work so far did not investigate neurophysiological or psychological causes for the changes of the brain patterns.

In this paper, we present a systematic quantitative study of data for multiple subjects recorded during offline and online sessions. The methods for analysis of the data and visualization thereof are applicable in general, even beyond BCI research, and provide a closer insight into the structure of the—global and local—changes in the EEG data. We study the distributions of task-relevant EEG features and provide evidence of changes both in the transition from offline to online settings and in the course of a single online session. We show that the former change can be interpreted as a shift of the data in feature space, due to the different background activity of the brain during the online feedback task (see section 3.2).

In the second part of our study, we propose adaptive classification techniques for use in BCIs with CSP (common spatial patterns)-based features. We designed our schemes (see section 4) in order to gain a quantitative understanding of the change in performance, and thereby suggest remedial schemes for improving online BCI performance. We applied our adaptive techniques to a variety of datasets collected from five subjects during online BCI control.

Our results demonstrate that although instabilities in BCI control can be encountered throughout the experiment, the major detrimental influence on the classification performance is caused by the initial shift from training to the test scenario. Hence, simple techniques that relearn only part of the classifier can overcome this change and can thus significantly improve BCI control.

This study focuses on a feature space that is a low-dimensional projection of 128-channel EEG data computed by the CSP algorithm [12, 14]. However, the methods of analysis, measurement and visualization, as well as the questions regarding adaptivity addressed in this paper, are widely applicable and should serve as useful tools in studying adaptivity in the BCI context.

2. Data from offline and online experiments

2.1. The Berlin BCI

The Berlin brain–computer interface was developed in cooperation with the data analysis group at Fraunhofer FIRST and neurologists of Campus Benjamin Franklin, Charité Berlin (cf http://www.bbci.de). We use event-related (de-)synchronization (ERD/ERS) features [26] in EEG signals related to hand and foot imagery as classes for control. These phenomena are well-studied and consistently reproducible features in EEG recordings and are used as the basis of a number of BCI systems (e.g. [8, 13]). Our EEG-to-control-signal translation algorithm consists of two parts. We first use a supervised feature selection algorithm called common spatial patterns (CSP) [2, 12–14] that dramatically reduces the dimensionality of the data from about 128 channels to 2–6 CSP projections. In this algorithm, the covariance matrices of the two different classes are diagonalized simultaneously in order to find the subspaces which have the largest variances for one class while minimizing the variance for the other class (for extensions see e.g. [8–10, 18]). Thus, the chosen dimensions of the feature space are those that contain maximal discriminative information in terms of amplitude modulations. We then perform further data reduction by using the log variance of a temporal window of data from each CSP channel, i.e., a single feature per channel remains. The power of this feature selection and data reduction scheme is demonstrated by the high separability of the resulting classes of data. In this setup, we use linear discriminant analysis (LDA) to separate data points with high accuracy into classes in the low-dimensional feature space. Note that more elaborate paradigms or other feature extraction techniques may require

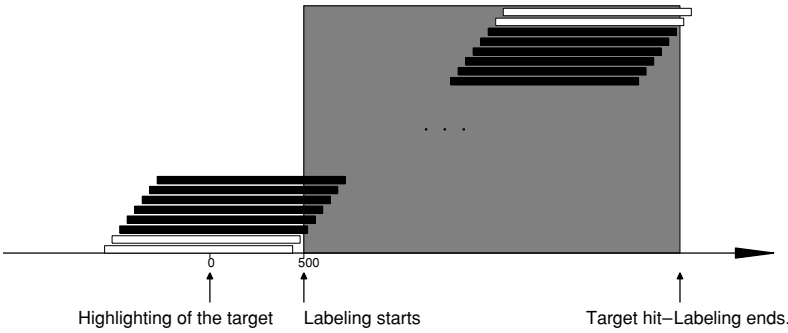

Figure 1. In the feedback session, sliding windows were used for classification. For adaptation and evaluation, we select the windows (here colored black) between releasing the cursor and the end of the trial. See the text for details.

the use of non-linear classifiers (cf [21, 23, 25, 31]). Previous work [3, 5, 8] has reported on the efficacy of our classification scheme. Other physiological paradigms implemented in the BBCI focus on the use of the lateralized readiness potential [4, 15, 16] and the combination of this feature with ERD/ERS [7, 8].

2.2. Experimental protocol

We investigate data from a BCI study consisting of experiments with six subjects[4]. For one subject, no effective separation of brain pattern distributions could be achieved. Thus, no feedback sessions were recorded and the dataset is left out in this investigation. All experiments were conducted at Fraunhofer FIRST in cooperation with the Department of Neurology of the Charité Berlin. The subjects were seated in a comfortable chair with arms lying relaxed on the armrests. In the calibration measurement (also called training or offline session), every 5.5 ($\pm$0.25) s one of three different visual stimuli was presented, indicating the motor imagery task the subject should perform for 3.5 s. The imagery tasks investigated were movements of the left hand (l), the right hand (r) and the right foot (f). Brain activity was recorded from the scalp with multi-channel EEG amplifiers using 128 channels. Besides EEG channels, we recorded the electromyograms (EMG) from both forearms and the right leg as well as horizontal and vertical electrooculograms (EOG) from the eyes. The EMG and EOG channels were exclusively used for monitoring to make sure that the subjects performed no real limb or eye movements correlated with the mental tasks that could directly (artifacts) or indirectly (afferent signals from muscles and joint receptors) be reflected in the EEG channels and thus be detected by the classifier, which operates on the EEG signals only. One hundred and forty trials were recorded for each class. These data were then used to train a classifier for the two best discriminable classes, using the above classification scheme (see [3, 8]). Subsequently, two feedback sessions were recorded where two targets were placed, one at each side of the screen. A 1 s window of data was used to estimate the features, which were classified over overlapping windows every 40 ms (see figure 1).

[4] Three of the authors participated as subjects in the experiments.

The continuous output from the classifier was then used to move the cursor either in a position-controlled (i.e., the scaled classifier output maps directly to the horizontal position on the screen) or in a rate-controlled manner (i.e., the scaled classifier output was used to move the cursor by a small amount in the chosen direction). During each trial, one of the targets was highlighted and the subject attempted to navigate the cursor into the target. Each trial lasted until the subject hit one of the two targets, and as a result the trials were of varying lengths. In a third experimental session, three rectangular targets were located at the bottom of the screen. A cursor was moving downwards with constant speed, while its horizontal position was controlled by the classifier output. Again, one of the targets was highlighted and the subject was instructed to try to hit the target with the cursor. The feedback sessions were recorded in a series of runs of 28 trials each, with short breaks in between runs.

2.3. Analyzing data from online sessions

Since the online sessions were controlled (i.e., the subject was directed to hit a certain target), we can use this information to label the data collected during an online session. When analyzing the data offline, as we will in the following, we have all the labels from the feedback experiment at our disposal and can use them in our evaluations. For labeling the data from an online session, we take the signals from the start of each trial until its completion and process the signals in a manner similar to the online scenario, i.e., compute features on overlapping windows of the same size and overlap as used in the online protocol. These data points are labeled according to the appropriate target class. Each trial may yield a different number of labeled data points since the trials were of varying length. When using the recorded data for testing various classification schemes, we always assign samples coming from one trial either all to the training or all to the test set.

It should be noted that in a realistic BCI scenario the labels of ongoing trials may not always be available; however, in some applications such as the use of a speller for communicating words, it is possible to estimate the labels *a posteriori* with high probability. Also, it is important to remember that the data were collected when the subject was using one particular classifier (the optimal classifier for the

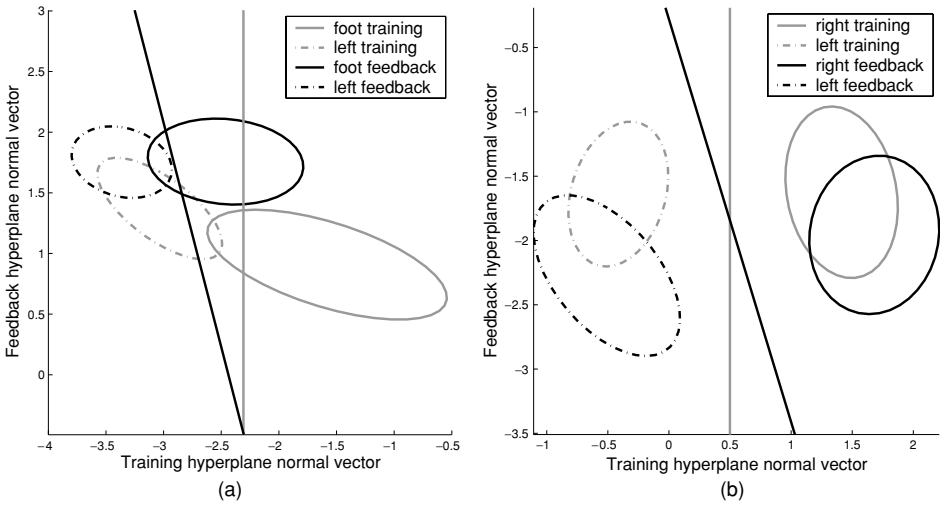

Figure 2. Changes in the optimal classifier from training to test. The figure shows, for subjects av and ay, the optimal hyperplane separating the training data classes (offline) and the test data classes (online). Also shown are the mean and covariance of the respective data distributions. In the case of subject av (a), the original classifier would perform very poorly, whereas for subject ay, as indicated in (b), the change is less severe.

training data, along with any manual adjustments to it) for BCI control. Clearly, the present offline analysis results will subsequently have to be further investigated in future experiments with online control.

3. Changes in the distributions of EEG features

In this section, we examine the changes in performance of the subjects using a variety of measures and new ideas for visualization that help us to characterize the type and degree of changes seen in EEG features used for BCI classification. In our study, we use the feature projections chosen by the CSP algorithm (see the previous section). We also strive to link these findings to possible neurophysiological changes that may cause these observed changes. We use two methods of visualizing the data: (1) by fitting a Gaussian distribution[5] on the data over an entire session (or over short-term windows), and (2) by examining the optimal separating hyperplane computed using an LDA classifier on the chosen data.

3.1. Differences from calibration measurement to feedback

Figure 2 shows a comparison between training data collected offline and the test data recorded during a subsequent online session. The figure shows, for two subjects, the hyperplanes of the classifiers computed on the training and test data, respectively, along with the means and covariances of the data points from each class. For ease of visualization, we have projected the data onto two specifically chosen dimensions (see the appendix) containing maximal information. We see from figure 2(a) that for subject av the test data distributions look very different from the training data, and in fact, the original classifier would perform quite poorly in the online scenario. This is not always the case, though—for example, in subject ay (figure 2(b)), while the test distributions are different from the training data, the impact of this change on online performance is less severe.

Table 1. Measuring the changes in the optimal classifier for offline and online distributions. These are the changes necessary for the classifier to perform optimally on feedback data, for every experiment in this study. Part (a) shows the ratio between the optimal shift for correcting the bias and the distance between class means. Part (b) shows the angle between the old hyperplane (calculated from the offline data) and the optimal hyperplane for the feedback data.

	Subject				
	al	aw	av	ay	aa
		(a)			
Shift/distance	0.11	0.80	0.83	0.07	0.26
	0.12	0.94	0.56	0.09	0.26
	0.01	0.82	0.61	0.06	0.60
		(b)			
Angles (°)	13.2	26.6	15.1	15.1	9.5
	9.7	20.6	28.7	17.7	6.7
	36.2	45.4	4.2	40.5	13.3

In order to examine this change more closely across all online datasets, we consider the following two possibilities for modifying the training classifier hyperplane: (1) shift the original classifier's hyperplane parallel to itself[6] in order to get the best performance in the online setting, and (2) in addition, rotate the hyperplane to further improve performance on the online data. We call these two methods REBIAS and RETRAIN. Table 1 summarizes the shift and angle required for optimal performance on each online dataset.

In order to understand the scale of the optimal shift, table 1(a) shows this shift as a fraction of the training data's class mean distance from the training classifier's hyperplane.

[5] On the plausibility of the assumption of Gaussian distributions in EEG data, see e.g. [4] and also the discussion in [23].

[6] This can be implemented, e.g., by simply adding a *bias* to the classifier output.

Table 2. Estimating the expected gain in classification when adapting the separation as calculated from the offline distributions to the online distributions. Any linear decision boundary between two normally distributed random variables misclassifies a certain quantile of both distributions. Here we compared the expected error quantiles for the optimal decision boundary for the training set to the decision boundary for the feedback sessions, when applied to the estimated distributions of the feedback data. Part (a) reflects the gain when only readapting the bias, and part (b) shows the improvement when the complete decision boundary is recalculated.

	Subject				
	al	aw	av	ay	aa
	(a)				
REBIAS/ORIG	0.93	0.79	0.67	1.00	0.97
	0.89	0.74	0.75	0.95	0.93
	1.00	0.75	0.80	0.99	0.82
	(b)				
RETRAIN/REBIAS	0.98	0.99	0.99	0.98	0.98
	0.98	0.99	0.94	0.71	0.98
	0.72	0.87	1.00	0.73	0.97

Note that in some cases the optimal shift is comparable to the distance of one class mean to the decision boundary. This shows that an adaptation of the bias would be necessary for correct classification. Table 1(b) shows the angle between training and test classifiers' hyperplanes on each dataset. In most cases, the angle does not change substantially. Table 2 provides an interpretation of these classifier changes in terms of their impact on classifier performance.

We show the ratios of estimated error quantiles for the training decision boundary, the bias-adapted decision boundary (table 2(a)) and the readapted decision boundary (table 2(b)). It is evident that the adaptation of the bias results in a significantly lower error quantile estimate, which confirms the findings in table 1, whereas an additional adaptation of the angle only gives a comparatively small gain.

3.2. Explaining the shift in data distributions

Figure 2 and table 1 together indicate that the primary difference between the offline and online situations is a *shift* of the data distributions for both classes in feature space, while not significantly changing their orientation. The source of this shift can be deduced from the spatial distributions of the band power on the scalp for the training and feedback situations.

As mentioned in section 2, we use the CSP algorithm for feature extraction and the classifiers are trained on these features under the assumption that the spatial distribution of these activation patterns remains fairly stable during feedback.

This assumption can be verified in figure 3 which displays task-specific brain patterns during offline versus online session for one representative subject. The scalps with red resp. blue circles show band power during left hand resp. right foot motor imagery, calculated from offline (upper row) and online (middle row) sessions. In the plots of the offline session, no systematic difference between the mental states can be seen, since the maps are dominated by a strong parietal α rhythm. Nevertheless, the map of r-values (see the appendix) reveals a difference focused over sensorimotor cortices. The parietal α rhythm is much less pronounced during the online session (middle row), resulting in a very strong difference between offline and online topographies, see the r-value maps in the lower row. In spite of this strong difference, the relevant difference between the tasks is qualitatively very similar in the offline and online settings (see the r-value maps in the right column). The topography of the difference between offline and online situations suggests that in the former case a strong parietal α rhythm (idle rhythm of the visual cortex) is present due to the decreased visual input during the calibration measurement, while that rhythm activity is decreased in online operation due to the increased demand for visual processing. The power spectra shown in figure 4 corroborate this assumption, since at parietal locations there is an increase in the power of the lower α band (just below 10 Hz).

Thus, there is a difference in *background activity* of the brain in offline and feedback settings. This difference also strongly influences the CSP features chosen for classification, cf section 3.3. This shift in feature space implies that the old classifier will perform poorly in these new settings without classifier adaptation.

3.3. Changes in EEG features during online sessions

We now examine the performance of subjects in the course of a single online session.

At each point of an online session, we consider a window for each class containing all data points from the last ten trials of that class. These data points can be used to get a *local estimate* of the density of each class at that point in time. We fit a Gaussian distribution to these local windows of data, as well as an *overall* density estimate for the entire online session.

Figure 5 shows for subject av the Kullback–Leibler (KL) divergence between the local density estimate for each class and the overall density estimate of that class over time. Since these curves alone do not provide information about classifiability of the data, we also show sample visualizations of data from certain time intervals, along with the classifier hyperplane. We see that the data distribution for the foot class changes over the course of the experiment, and the KL divergence curve reflects these changes.

The subject's success in controlling the BCI was fairly varying, and the short period of time where the KL divergence for the foot class is very high corresponds to a period when the subject gained better control over the BCI. This can also be inferred from the visualizations of the distributions presented in the lower portion of the figure. A point to be noted is that breaks between runs may also affect performance. For example, one of the breaks coincided with the end of the phase with good performance—it is possible that upon resuming the experiment the subject was unable to regain the control acquired in the previous phase. For a closer look, we plot the data distributions from each uninterrupted run in figure 6. A further study consisting of new long-term experiments is needed for separating such gradual and sudden changes and

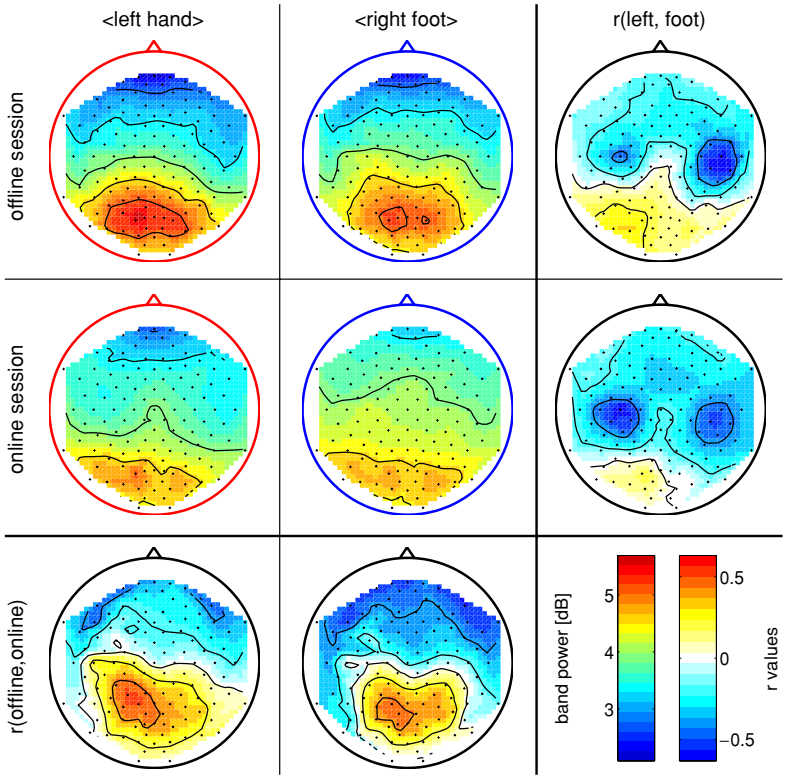

Figure 3. This figure shows the task-specific brain patterns and how they differ between offline and online sessions. The upper left 2×2 matrix of scalps displays topographic scalp maps (view from the top, nose up) of band power (broadband 7–30 Hz as used for calculating the CSP features in this subject). Maps are calculated from the offline session (upper row) resp. online session (middle row) separately for motor imagery of the left hand (left column) resp. of the right foot (middle column). Maps in the right column show the r-values of the difference between the tasks, and maps in the lower row show the r-values of the difference between offline and online sessions. While there is a huge and systematic difference between brain activity during offline and online sessions, the significant difference between the tasks stays fairly stable when going from offline to online operation (compare the r-value maps in the right column).

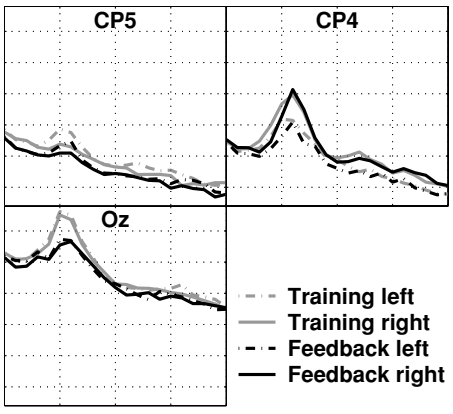

Figure 4. This figure shows the spectra in the frequency range 5–25 Hz both in training and in feedback, for the two classes separately. The amplitudes are in the range 22–54 dB.

providing further insight on the highly individual lapses of performance, but is beyond the scope of this paper. It is, however, clear and quantified in the present paper that the user's performance over a short period of time (about 30 min) can show considerable changes.

4. Adaptive classification

We have shown qualitative and quantitative evidence indicating non-stationarity in the BCI classification problem; however, two questions remain unanswered so far: (a) what is the impact of this non-stationary behavior on performance in a feedback setting? (b) What remedial measures can we use to address the non-stationary behavior of EEG-related features? In this section, we propose a range of techniques that aim to quantify the nature and impact of non-stationarity on performance, and thereby suggest adaptive methods for improving online control. Accordingly, we describe the various classifiers that we compare and the rationale behind each choice, and subsequently discuss their applicability in an online scenario.

4.1. Adaptive methods

The adaptive classification methods investigated are as follows:

- *ORIG*: this is the unmodified classifier trained on data from the offline scenario and serves as a baseline.
- *REBIAS*: we use the continuous output of the unmodified classifier and *shift* the output by an amount that would minimize the error on the labeled feedback data.

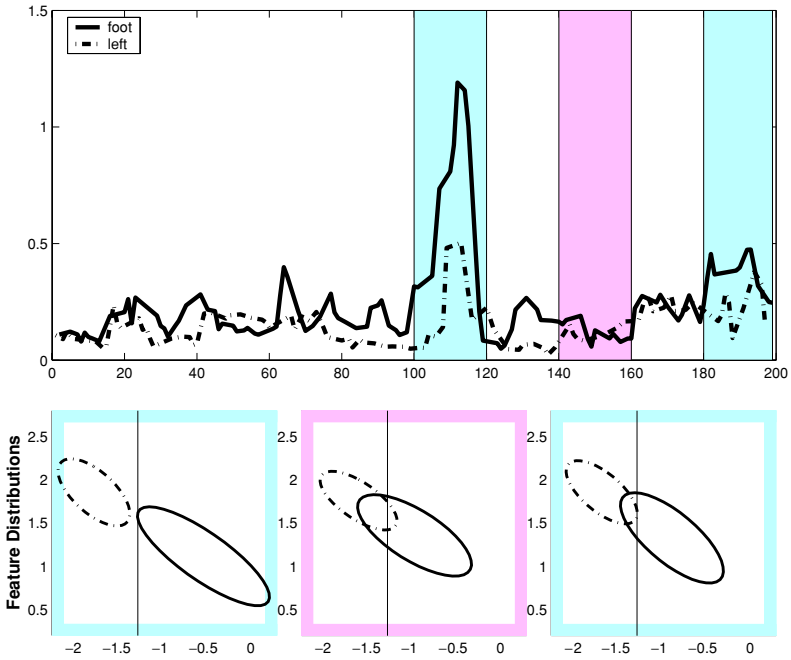

Figure 5. This figure shows the change of the Kullback–Leibler divergence during the feedback session. The corresponding feature distributions are displayed below for the shaded intervals. The data are projected on the plane spanned by the normal vector of the optimal separating hyperplane for the feedback and the largest PCA component of the feedback data (see the appendix).

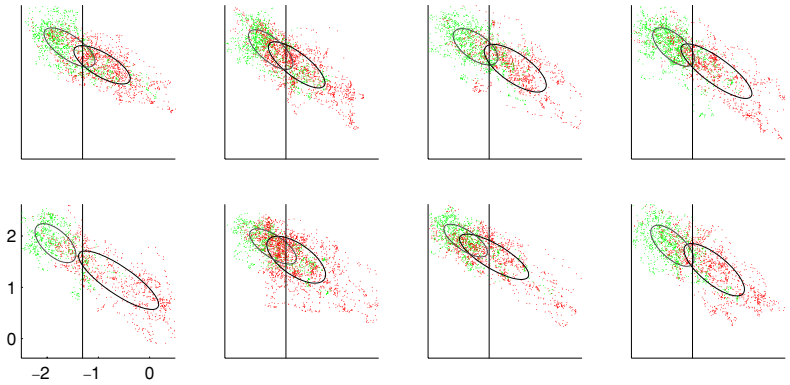

Figure 6. The single plots in this figure represent the development of the feature distributions for subject av throughout one feedback experiment, windows representing each run (consisting of 28 trials each). The data are projected on the feature subspace spanned by the optimal hyperplane and the largest PCA component (see the appendix).

- *RETRAIN*: we use the features as chosen from the offline scenario, but retrain the LDA classifier to choose the hyperplane that minimizes the error on labeled feedback data.
- *RECSP*: we completely ignore the offline training data and perform CSP feature selection and classification training solely on the feedback data.

The schemes are listed in increasing order of change to the classifier and correspond to different assumptions on the degree of difference between offline and online data. In addition, we have the option of using (1) *all* the labeled online data up to the current point (cumulative), (2) only a window over the immediate past (moving), or (3) only an initial window of data from each session (initial). Each choice corresponds to different assumptions of the volatility of the online classification problem. We thus have C-REBIAS[7], C-RETRAIN and C-RECSP, W-REBIAS, W-RETRAIN and W-RECSP, and I-REBIAS, I-RETRAIN and I-RECSP, respectively, for the three cases considered.

4.2. Performance against non-adaptive classifiers

Figure 7(a) compares the classification error of each adaptive method with the non-adaptive ORIG classifier. The adaptive classifiers were trained on a window of 60 s length. This

[7] C- denotes cumulative, W- denotes fixed window sizes and I- denotes use of only the initial segment of the session.

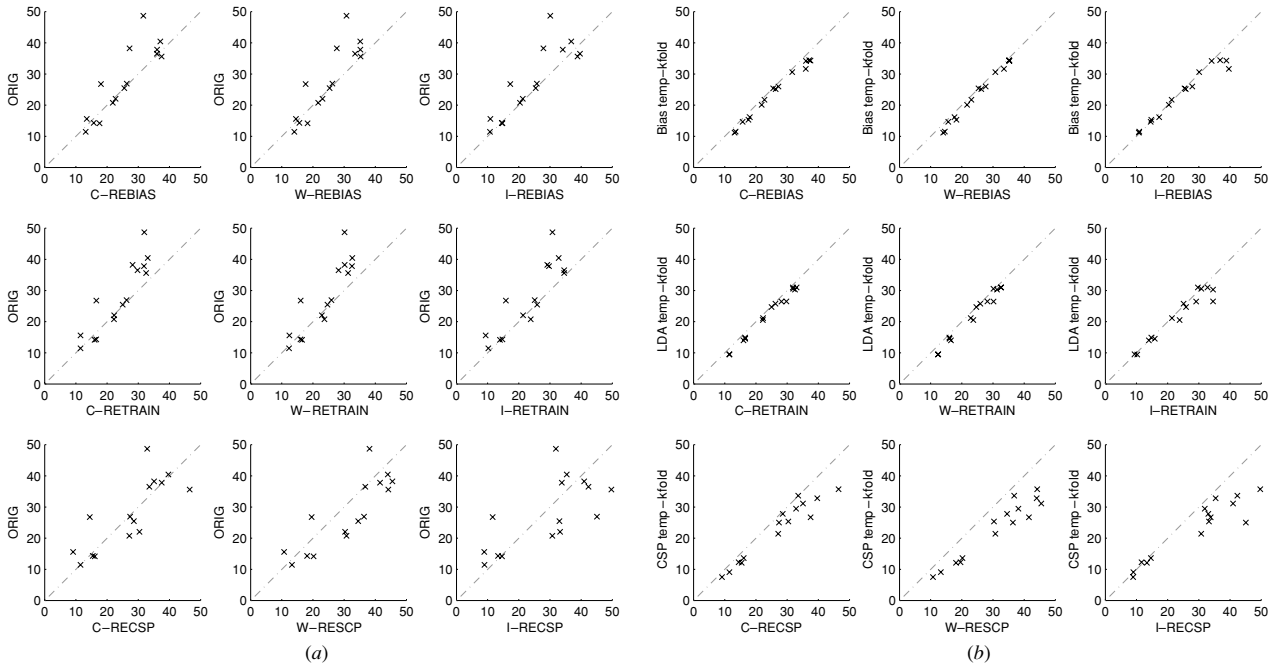

Figure 7. Comparison of various adaptive classification methods on data recorded from online sessions. Each subplot is a scatter plot, with the error rate of a reference method on the *y*-axis and the error rate of the method of investigation on the *x*-axis. The performance of the latter is better for those data points that lie over the diagonal. Error rates are given in percentage. (a) All the REBIAS and RETRAIN variants clearly outperform the unmodified classifier trained on the offline data. (b) The adaptive methods are compared against a theoretical cross-validation error baseline that uses labels of future data points in the online session. See the text for more details.

was also the shortest (i.e., first) window of the cumulative classifiers.

Each row presents the three different possibilities for training data, and each column presents the three adaptation methods considered. Inspecting each column, we see that the schemes REBIAS and RETRAIN clearly outperform the ORIG classifier, since most of the classification errors on the feedback data decrease. RECSP, on the other hand, does not improve performance. A possible reason for this is the small training sample size, a question we will revisit in the next section. Further, when examining each row, we see that the I- methods perform better than the W- and C- methods, indicating that the I- methods are more stable than the C- and W- methods.

Also, on examining all nine plots in figure 7(a), we see that the I-REBIAS method is comparable to all the other schemes. This is a very useful result because the I-REBIAS method is a lightweight adjustment that only requires a *short initial calibration period* and is thus relatively non-intrusive. Thus, figure 7(a) shows that adaptive methods can indeed improve performance, even with simple adaptive schemes.

4.3. Performance against baseline cross-validation error

We now examine the following question regarding the online BCI scenario: how non-stationary is the data distribution within the online sessions? For each method, we define an idealized baseline scenario where the method can access the data and labels of both past and future from an online session. We then compare the temporal[8] *k*-fold cross-validation error of the method in this baseline scenario to the method trained only on data from the past (as in the previous experiment).

This choice of baseline is aimed at examining whether each method suffers from having 'too much' training data, or too little data. For example, if the classification problem were highly non-stationary, we would then expect the windowed methods to outperform the baseline, since they can adapt to local changes. If the data are fairly stable across an online session, then the baseline cross-validation error would be lower, since it has more training data.

Figure 7(b) shows the results of this comparison. We can make the following inferences from the figure: first, the baseline is better in almost all cases, indicating that the adaptive methods have insufficient data. This is especially true for the RECSP algorithms and is clearly because of the very high dimensional data they deal with. Second, the REBIAS methods do not benefit very much by the addition of more data, and the I-REBIAS error is comparable to the temporal *k*-fold error on REBIAS. This does not necessarily mean that there are no dynamic changes in the data; in fact, in section 3.3 we see that the data distributions move around considerably. Instead, these results indicate that within the constraints of the chosen feature space and the adaptive algorithm, more training data will not help. Thus, the changes in the data are more in the nature of phases where the separability of the data is poor. The positive result from this experiment is that the performance of

[8] That is, the data are divided into *k* contiguous blocks in order to prevent overfitting.

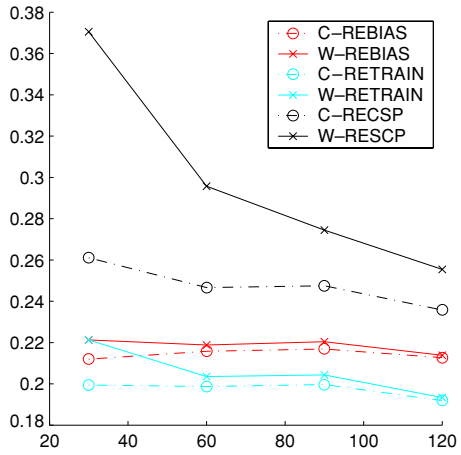

Figure 8. Influence of parameters on the adaptive classification results. This figure shows the average error across all sessions and subjects as a function of the window of data points (in seconds) used for the windowed classification methods. For the C- classifiers, this indicates the size of the first training window.

the REBIAS algorithm, using only an initial window of data, is comparable to the ideal cross-validation baseline error for the REBIAS algorithm.

4.4. Increasing available training data

We now examine whether our choice of feature space is a factor in the performance of our classification algorithm. Figure 8 shows the error averaged across subjects for each dynamic version of the adaptive algorithms (i.e., the C- and W- methods), as a function of the data window used for training. The figure confirms that the RECSP methods indeed improve on addition of training data; however, they are still considerably worse than the best performing algorithm. Our experiments were not sufficiently long to examine whether, with sufficient data, the RECSP algorithms can be competitive.

5. Discussion

A proposal for adaptive algorithms in BCI has to address the following issues: (a) the need for adaptivity, (b) the possible sources of the change in the data, (c) an adaptive scheme that can demonstrably improve performance over the non-adaptive baseline algorithm, and (d) the impact of the adaptive algorithm on the subject who is trying to use the BCI for control.

We have shown that an important factor affecting online BCI performance is the neurophysiological change to the mental state of the subjects (as described in section 3.3) between the offline and online settings. Our results show both that in a CSP-based BCI system adaptive methods are necessary and that simple adaptive schemes can significantly improve performance. The adaptive method we recommend specifically addresses the change from training data to online performance and is in the form of a small one-time bias adjustment. As a result, we do not risk confusing the user with a continually changing interface, or overfitting data during the subject's familiarization phase.

Our results also indicate that this one-time adjustment is in fact sufficient for the time periods we have considered (up to 1 h of continuous use). The success of the simple adaptive schemes is mainly due to the effectiveness of our offline feature selection scheme and the fact that the basic neurophysiological processes used for control are similar in the offline and feedback scenarios.

While changes in performance and feature distributions do occur during online sessions (see section 3.3), our classification results indicate that on average they do not have a significant effect on performance. It remains unclear at this point whether these changes can be affected by a different choice of feature space or the use of additional features; however, a complete relearning of the feature selection is impractical due to the need for large amounts of labeled data. Our planned studies of longer term BCI operation aim to shed further light on the exact nature of the changes during an online setting.

6. Conclusion

EEG-based brain–computer interfaces frequently have to deal with a decrease in performance when going from offline calibration sessions to online operation of the BCI. One explanation for this decrease is that bad model selection strategies have resulted in overly complex classification models that overfit the EEG data [23]. The current work has clearly shown that an alternative reason for failure should also be considered: non-stationarities in the EEG statistics. The subject's brain processes during feedback can cause the distributions to wander astray on a very local time scale. This could in principle make classification a difficult problem, perhaps necessitating special statistical modeling that takes into account covariate shifts [29] or even more sophisticated techniques such as transductive inference [30]. However, the successful adaptive methods investigated in this work turn out to be surprisingly simple: a bias adaptation in combination with an offline feature selection scheme significantly increases BCI performance. We clearly demonstrated that the strongest source of non-stationarity stems from the difference between calibration and feedback sessions, whereas during the feedback session the statistics seems rather stable on the scale of up to an hour (depending on the subject). So a practical recommendation of this study is (1) to correct for the bias between calibration and feedback sessions, and (2) either to incorporate intermediate corrections every half hour with a short 2–3 min calibration or to adapt the bias when changes of the statistics, say due to fatigue, are observed. Future research will explore the use of transductive methods and dedicated statistical tests to detect and address non-stationarities automatically.

Acknowledgments

We thank G Dornhege, A Schwaighofer, F Meinecke, S Harmeling and G Curio for helpful discussions. The work

of MK, BB and KRM was supported in part by grants of the *Bundesministerium für Bildung und Forschung* (BMBF), FKZ 01 IBE 01A/B, by the *Deutsche Forschungsgemeinschaft* (DFG), FOR 375/B1 and MU 987/1-1, and by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views. Rajesh P N Rao was supported by NSF grant 130705 and the Packard Foundation. We thank the anonymous reviewers for their comments that helped us to improve the quality of this manuscript.

Appendix

A.1. Feature distribution projections

The lower part of figure 5 shows local estimates of the distributions of both classes during one feedback session. We first calculated the classifier which is optimal for the feedback session and the largest PCA component w_{PCA} of the features. The x-axis shows the projection of the data on normal vector w_{FB} of that hyperplane of the feature space corresponding to the decision boundary of the classifier. The other dimension is chosen orthogonally to w_{FB}, such that w_{PCA} is contained in this two-dimensional subspace. It is a property of this display mode that the relative location of the distributions to the hyperplane can be seen by orthogonal projection, and the dimension with the largest variance is contained in this plot, while preserving the angles of the original space.

Figure 2 is generated similarly, only the dimensions used here are the normal w_{TR} of the original classifier as obtained from the training session and the normal w_{TR} from the feedback classifier hyperplane (as above). The black and gray lines denote the intersections of the decision boundaries of the classifiers with the subspace which is shown here. Also in this case, the projection preserves angles.

A.2. Bi-serial correlation coefficients

In figure 3, we show the r-values r_{ch} of the band-power values $\mathrm{fv}_{\mathrm{ch}}$ in each channel ch. The bi-serial correlation coefficient r measures how much information one feature carries about the labels. It is computed in the following way:

$$r_{\mathrm{ch}} = \frac{(\mu_1 - \mu_2)\sqrt{\#\mathrm{cl}_1 \#\mathrm{cl}_2}}{\sqrt{\mathrm{var}(\mathrm{fv}_{\mathrm{ch}})}(\#\mathrm{cl}_1 + \#\mathrm{cl}_2)},$$

where μ_i is the class-specific mean of $\mathrm{fv}_{\mathrm{ch}}$ and $\#\mathrm{cl}_i$ denotes the number of trials for class $i \in \{1, 2\}$.

A.3. Kullback–Leibler distance

The Kullback–Leibler distance (or Kullback–Leibler divergence) of the probability distributions P and Q is defined by

$$\mathrm{KL}(P, Q) := \int p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x.$$

For two n-dimensional random variables X_1, X_2 with $X_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $X_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$, this amounts to

$$\begin{aligned}
&\mathrm{KL}\big(P_{X_1}, P_{X_2}\big) \\
&= -\tfrac{1}{2}\big[\log\big(\big|\Sigma_1 \Sigma_2^{-1}\big|\big) + E(X_1 - \mu_1)^t \Sigma_1^{-1}(X_1 - \mu_1) \\
&\quad - E(X_1 - \mu_2)^t \Sigma_2^{-1}(X_1 - \mu_2)\big] \\
&= -\tfrac{1}{2}\big[\log\big(\big|\Sigma_1 \Sigma_2^{-1}\big|\big) + \mathrm{trace}\big(E(X_1 - \mu_1)(X_1 - \mu_1)^t \Sigma_1^{-1}\big) \\
&\quad - \mathrm{trace}\big(E(X_1 - \mu_1)(X_1 - \mu_1)^t \Sigma_2^{-1}\big) \\
&\quad - (\mu_2 - \mu_1)^t \Sigma_2^{-1}(\mu_2 - \mu_1)\big] \\
&= -\tfrac{1}{2}\big[\log\big(\big|\Sigma_1 \Sigma_2^{-1}\big|\big) + \mathrm{trace}\big(I - \Sigma_1 \Sigma_2^{-1}\big) \\
&\quad - (\mu_2 - \mu_1)^t \Sigma_2^{-1}(\mu_2 - \mu_1)\big],
\end{aligned}$$

where I denotes the n-dimensional identity matrix.

In figure 5, we estimated the overall feedback densities of both classes on all trials of the feedback session and displayed their Kullback–Leibler divergence to the local estimates of the densities, which are obtained by averaging over the features from the last ten trials of each class.

References

[1] Birbaumer N, Ghanayim N, Hinterberger T, Iversen I, Kotchoubey B, Kübler A, Perelmouter J, Taub E and Flor H 1999 A spelling device for the paralysed *Nature* **398** 297–8

[2] Blanchard G and Blankertz B 2004 BCI competition 2003–data set IIa: spatial patterns of self-controlled brain rhythm modulations *IEEE Trans. Biomed. Eng.* **51** 1062–6

[3] Blankertz B, Dornhege G, Krauledat M, Müller K-R and Curio G 2005 The Berlin brain–computer interface: report from the feedback sessions *Technical Report* 1 (Fraunhofer FIRST) http://ida.first.fraunhofer.de/publications/BlaDorKraMueCur05.pdf

[4] Blankertz B, Dornhege G, Schäfer C, Krepki R, Kohlmorgen J, Müller K-R, Kunzmann V, Losch F and Curio G 2003 Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis *IEEE Trans. Neural Syst. Rehabil. Eng.* **11** 127–31

[5] Blankertz B *et al* 2004 The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials *IEEE Trans. Biomed. Eng.* **51** 1044–51

[6] Curran E A and Stokes M J 2003 Learning to control brain activity: a review of the production and control of EEG components for driving brain–computer interface (BCI) systems *Brain Cogn.* **51** 326–36

[7] Dornhege G, Blankertz B, Curio G and Müller K-R 2003 Combining features for BCI *Advances in Neural Information Processing Systems (NIPS 02)* vol 15, ed S Becker, S Thrun and K Obermayer, pp 1115–22

[8] Dornhege G, Blankertz B, Curio G and Müller K-R 2004 Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms *IEEE Trans. Biomed. Eng.* **51** 993–1002

[9] Dornhege G, Blankertz B, Curio G and Müller K-R 2004 Increase information transfer rates in BCI by CSP extension to multi-class *Advances in Neural Information Processing Systems* vol 16, ed S Thrun, L Saul and B Schölkopf (Cambridge, MA: MIT Press) pp 733–40

[10] Dornhege G, Blankertz B, Krauledat M, Losch F, Curio G and Müller K-R 2006 Optimizing spatio-temporal filters for improving brain–computer interfacing *Advances in Neural Information Processing Systems (NIPS 05)* vol 18, at press

[11] Elbert T, Rockstroh B, Lutzenberger W and Birbaumer N 1980 Biofeedback of slow cortical potentials: I *Electroencephalogr. Clin. Neurophysiol.* **48** 293–301
[12] Fukunaga K 1990 *Introduction to Statistical Pattern Recognition* 2nd edn (Boston, MA: Academic)
[13] Guger C, Ramoser H and Pfurtscheller G 2000 Real-time EEG analysis with subject-specific spatial patterns for a brain–computer interface (BCI) *IEEE Trans. Neural Syst. Rehabil. Eng.* **8** 447–56
[14] Koles Z J and Soong A C K 1998 EEG source localization: implementing the spatio-temporal decomposition approach *Electroencephalogr. Clin. Neurophysiol.* **107** 343–52
[15] Krauledat M, Dornhege G, Blankertz B, Curio G and Müller K-R 2004 The Berlin brain–computer interface for rapid response *Biomed. Tech.* **49** 61–2
[16] Krauledat M, Dornhege G, Blankertz B, Losch F, Curio G and Müller K-R 2004 Improving speed and accuracy of brain–computer interfaces using readiness potential features *Proc. 26th Annual Int. Conf. IEEE EMBS on Biomedicine (San Francisco)*
[17] Kübler A, Kotchoubey B, Kaiser J, Wolpaw J and Birbaumer N 2001 Brain–computer communication: unlocking the locked in *Psychol. Bull.* **127** 358–75
[18] Lemm S, Blankertz B, Curio G and Müller K-R 2005 Spatio-spectral filters for improved classification of single trial EEG *IEEE Trans. Biomed. Eng.* **52** 1541–8
[19] McFarland D J, Miner L A, Vaughan T M and Wolpaw J R 2000 Mu and beta rhythm topographies during motor imagery and actual movements *Brain Topogr.* **12** 177–86
[20] Millán J D R 2004 On the need for on-line learning in brain–computer interfaces *Proc. Int. Joint Conf. on Neural Networks (Budapest, Hungary, July 2004)* (IDIAP-RR 03-30)
[21] Millán J D R, Mouriño J, Franzé M, Cinotti F, Varsta M, Heikkonen J and Babiloni F 2002 A local neural classifier for the recognition of EEG patterns associated to mental tasks *IEEE Trans. Neural Netw.* **13** 678–86
[22] Millán J D R, Renkens F, J M no and Gerstner W 2004 Non-invasive brain-actuated control of a mobile robot by human EEG *IEEE Trans. Biomed. Eng.* **51** 1026–33
[23] Müller K-R, Anderson C W and Birch G E 2003 Linear and non-linear methods for brain–computer interfaces *IEEE Trans. Neural Syst. Rehabil. Eng.* **11** 165–9
[24] Müller K-R, Krauledat M, Dornhege G, Curio G and Blankertz B 2004 Machine learning techniques for brain–computer interfaces *Biomed. Tech.* **49** 11–22
[25] Müller K-R, Mika S, Rätsch G, Tsuda K and Schölkopf B 2001 An introduction to kernel-based learning algorithms *IEEE Neural Netw.* **12** 181–201
[26] Pfurtscheller G and da Silva F H L 1999 Event-related EEG/MEG synchronization and desynchronization: basic principles *Clin. Neurophysiol.* **110** 1842–57
[27] Pfurtscheller G and Neuper C 1997 Motor imagery activates primary sensorimotor area in humans *Neurosci. Lett.* **239** 65–8
[28] Rockstroh B, Birbaumer N, Elbert T and Lutzenberger W 1984 Operant control of EEG and event-related and slow brain potentials *Biofeedback Self-Regul.* **9** 139–60
[29] Sugiyama M and Müller K-R 2006 Input-dependent estimation of generalization error under covariate shift *Statistics and Decisions* at press (http://www.cs.titech.ac.jp/tr/reports/2005/TR05-0001.pdf)
[30] Vapnik V 1998 *Statistical Learning Theory* (New York: Wiley)
[31] Vidaurre C, Schlögl A, Cabeza R and Pfurtscheller G 2004 About adaptive classifiers for brain–computer interfaces *Biomed. Tech.* **49** 85–6
[32] Wolpaw J R, Birbaumer N, McFarland D J, Pfurtscheller G and Vaughan T M 2002 Brain–computer interfaces for communication and control *Clin. Neurophysiol.* **113** 767–91
[33] Wolpaw J R and McFarland D J 2004 Control of a two-dimensional movement signal by a noninvasive brain–computer interface in humans *Proc. Natl Acad. Sci. USA* **101** 17849–54