TB, KR, JNE/190440, 23/06/2005

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF NEURAL ENGINEERING

J. Neural Eng. **2** (2005) L1–L9 doi:10.1088/1741-2560/2/0/000

COMMUNICATION

Characterization of four-class motor imagery EEG data for the BCI-competition 2005

Alois Schlögl[1], **Felix Lee**[2], **Horst Bischof**[2] **and Gert Pfurtscheller**[2]

[1] Institute of Human-Computer Interfaces, University of Technology Graz, Krenngasse 37, A-8010 Graz, Austria
[2] Institute of Computer Vision and Graphics, University of Technology Graz, Inffeldgasse 16a, A-8010 Graz, Austria

E-mail: alois.schloegl@tugraz.at

Received 24 November 2004
Accepted for publication 10 June 2005
Published DD MMM 2005
Online at stacks.iop.org/JNE/2/L1

Abstract
To determine and compare the performance of different classifiers applied to four-class EEG data is the goal of the present letter. The EEG data were recorded with 60 electrodes from five subjects performing four different motor-imagery tasks. The EEG signal was modeled by an adaptive autoregressive (AAR) process whose parameters were extracted by Kalman filtering. By these AAR parameters four classifiers were obtained, namely minimum distance analysis (MDA)—for single-channel analysis, and linear discriminant analysis (LDA), *k*-nearest-neighbor classifier (*k*NN) as well as support vector machine (SVM) classifiers for multi-channel analysis. The performance of all four classifiers was quantified and evaluated by Cohen's kappa coefficient, an advantageous measure we introduced here to BCI research for the first time. The single-channel results gave rise to topographic maps that revealed the channels with the highest level of separability between classes for each subject. Our results of the multi-channel analysis indicate SVM as the most successful classifier, whereas *k*NN performed worst.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

The growing interest toward brain–computer interfaces (BCI) is most likely linked to one of their core functions, to establish a direct connection between the human brain and electronic or mechanical devices. Recently obtained results underlined that the application of BCI helps humans to restore their motor function and communication ability lost through injury or disease (Birbaumer *et al* 2003, Pfurtscheller *et al* 2003, Wolpaw *et al* 2002). In order to establish a successful BCI system, several key components have to be taken into account, in particular high-quality EEG recordings, subjects' motivation and involvement, most accurate and fast ways of signal analysis to discriminate and characterize different brain states reflected by the ongoing EEG.

In the face of the great variety of different methods within BCI research that are used to analyze and classify EEG signals, the comparison of these different approaches means a necessary evaluation of their potential impact. For this reason a general competition between several BCI research groups was initiated so that each participating group could prove the performance of their approaches. The start of such a competition means to make a certain data set available to all participants so that they are able to apply and adjust their methods. After each group has submitted their developed classification methods, their individual approaches

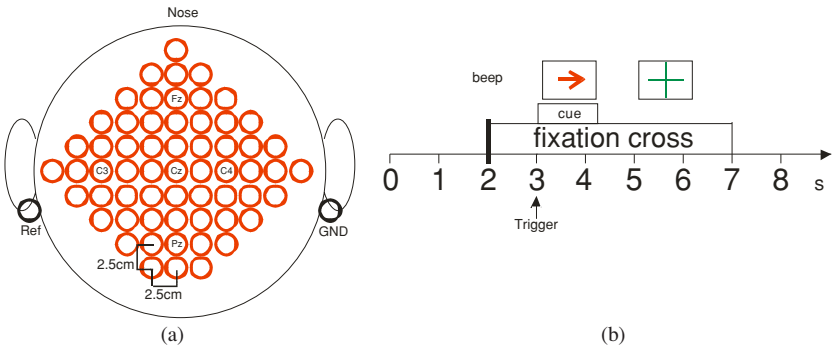

Figure 1. Position of EEG electrodes (a) and timing of the training paradigm (b).

are evaluated by applying them on a new, up to then (to the participants) unknown 'test' data-set. Finally, the submitted classification methods are ranked based on their performance. Such a competition is a fair comparison between different methods and provides important impetus to the development of BCI research. In the last BCI competition, held in 2003 (Blankertz *et al* 2003), the dataset from the Graz group contained two classes and requested continuous classification. The 60-channel data set from Graz for the 2005 BCI competition resulted from a four-class classification task. The current report addresses three main points:

1. Description of the data set (IIIa) made available for the BCI competition 2005.
2. Comparison of the several classification approaches: i.e. linear discriminant analysis (LDA), minimum distance analysis (MDA), k-nearest-neighbor (kNN) classifier and support vector machines (SVM).
3. Introduction of kappa as an advantageous criterion for BCI data classification with multiple classes.

2. Data acquisition, preprocessing and classification methods

2.1. Data acquisition

The data sets were recorded from five subjects, K3, K6, L1, P19 and Q5 by using a 64-channel Neuroscan EEG amplifier. The left mastoid served as reference and the right mastoid as ground. The EEG was sampled with 250 Hz and filtered between 1 and 50 Hz. A notch filter was enabled to suppress line noise. The 60 EEG channels recorded were placed according to the scheme in figure 1.

In fact our training paradigm consisted of a sequential repetition of cue-based trials. The subjects were seated in a relaxing chair with armrests and instructed to perform imagery movements prompted by a visual cue. Each trial started with an empty black screen; at time point $t = 2$ s a short beep tone was presented and a cross '+' appeared on the screen to raise the subject's attention. Then at second 3 ($t = 3$ s) an arrow appearing for 1.25 s pointed either to the left, right, upwards or downwards. Each position indicated by this arrow instructed the subject to imagine either a left hand, right hand, tongue or foot movement, respectively. The respective movement imagination should be performed until the cross disappeared at $t = 7$ s (see figure 1(a) and 2). The next trial started after a 3.24 s second resting period, while the EEG was continuously recorded. Each of the four types of cues was displayed ten times within each run in a randomized order. No feedback was provided to the subject. The data set recorded from subject K3 consisted of 9 runs, whereas the data set of K6 and L1 consisted of 6 runs each. The numbers of runs and trials recorded from each subject are summarized in table 1. For the sake of subsequent analysis the data of all runs were concatenated and converted to the GDF format. Three data sets (K3, K6 and L1) have been already used for the BCI competition 2005.

Table 1. Overview of the number of trials for each recording.

Subject	Number of runs	Number of trials	Left hand	Right hand	Foot	Tongue
K3	9	360	90	90	90	90
K6	6	240	60	60	60	60
L1	6	240	60	60	60	60
P19	6	240	60	60	60	60
Q5	8	320	80	80	80	80

2.2. Preprocessing: feature extraction by estimating AAR parameters

First, the raw EEG data were down-sampled from 250 Hz to 125 Hz. Second, to capture the spectral components (i.e. second-order statistics) of the EEG an AR process was constructed that modeled the recorded EEG signal. Mathematically, the AR model (of order p) is described by the following equation:

$$y_k = a_1 y_{k-1} + a_2 y_{k-2} + \cdots + a_p y_{k-p} + x_k = \vec{y}_{k-1}{}^T \vec{a} + x_k,$$

where a_i express the autoregressive parameters, y_i the observed sample values, x_i a zero mean white noise process; $\vec{a}$ is a vector of p AR parameters, and $\vec{y}_{k-1}$ is the vector of the past p sample values.

In order to consider the variation of the EEG spectrum over time, the autoregressive parameters have to change over time as well. Therefore, such kind of parameters are termed adaptive autoregressive (AAR) ones. The application of this kind of AR parameters seems particularly indicated in terms of online and real-time computations. As the calculation of the AR parameters is concerned Kalman filtering was the method of

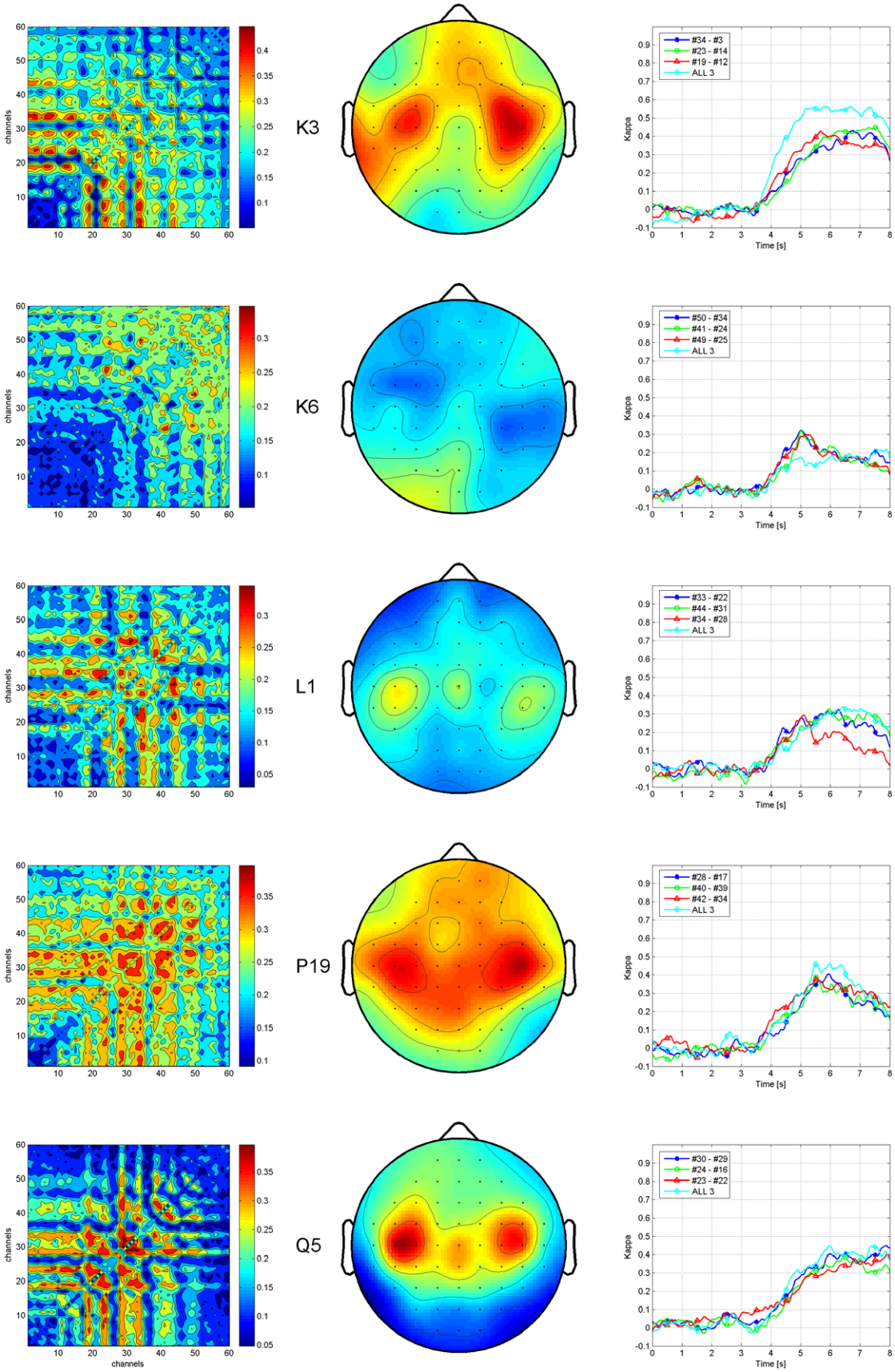

Figure 2. Single-channel analysis of subjects K3, K6, L1, P19, Q5 (from top to bottom): (a) the first column displays the maximum kappa for each monopolar (diagonal) and bipolar (off-diagonal) channels. (b) The topographic map displays the relative importance of each electrode. (c) The third column shows the time courses of the kappa coefficient for the three best single channels and the combination of these three channels. The time courses are smoothed with a triangular windowing function with a length of 0.2 s.

choice, for several reasons: first, it is an adaptive and, second, also a causal algorithm (i.e. using only data from the past) and third, Kalman filtering was shown to be an optimal estimator in terms of second-order statistics (Schlögl 2000). The update equations of the Kalman filter are defined as follows:

$$\begin{aligned} e_k &= y_k - \vec{y}_{k-1}{}^T \vec{a}_k \\ Q_k &= \vec{y}_{k-1}{}^T A_{k-1} \vec{y}_{k-1} + V_k \\ \vec{k}_k &= \frac{A_{k-1}\vec{y}_{k-1}}{Q_k} \\ \vec{a}_k &= \vec{a}_{k-1} + \vec{k}_k e_k \\ X_k &= A_{k-1} - \vec{k}_k \vec{y}_{k-1}{}^T A_{k-1} \\ A_k &= X_k + W_k. \end{aligned}$$

Only the covariance matrix W_k, the variance V_k and the initial values a_0 and A_0 need to be defined. In order to avoid initialization effects, the AAR estimation was calculated twice. The first run provides AAR estimates by using $\vec{a}_0 = [0, \ldots, 0]$, $A_0 = I_{p\times p}$, $V_k = 1 - UC$, $W_k = I \times UC \times \mathrm{trace}(A_{k-1})/p$ with $UC = 0.0055$ and $p = 3$. The AAR estimates a_t and the prediction error e_t of the first run are used to calculate the initial values for the second run, using $\vec{a}_0 = \mathrm{mean}\{\vec{a}_t\}$, $A_0 = \mathrm{cov}\{\vec{a}_t\}$, $V_k = \mathrm{var}\{e_t\}$, $W_k = \mathrm{cov}\{\Delta a_t\}$ with $\Delta\vec{a}_t = \vec{a}_t - \vec{a}_{t-1}$. In general, choosing the model order and the update coefficient means always a trade-off between adaptation speed and estimation accuracy (Schlögl 2000). Given that the optimization of parameters is beyond the scope of this study, we defined both criteria upon our previous experience.

2.3. Classification methods: linear discriminant analysis, minimum distance analysis and k-nearest-neighbor classifier

LDA is one of the most popular classification methods. The basic idea of LDA is to find the best discriminating projection direction so that the distance between the classes is maximized, while the distance within a class is minimized. LDA is simple, robust and can be used to produce a continuous BCI output in time as well as in amplitude (Schlögl *et al* 1997). The minimum distance analysis (MDA) was based upon a certain distance measure, the Mahalanobis distance, assuming for each class a Gaussian distribution with mean μ_c and covariance Σ_c. The mean μ_c and the covariance Σ_c define the multivariate normal probability density function, that corresponds to class c. Any point in the n-dimensional feature space can be associated with a certain distance to each class c.

The Mahalanobis distance $d_c(x)$ of point x with respect to the multivariate normal distribution $N(\mu_c, \Sigma_c)$ is defined by

$$d_c^2(x) = (x - \mu_c) \sum_c^{-1} (x - \mu_c)^T,$$

with mean μ_c and the covariance Σ_c estimated from the training samples of class c. Accordingly, for each point x in the n-dimensional feature space, we yield a distance to each class c and assign x to the class with the smallest distance. Using this spatial information, simple and robust statistical classifiers can be obtained, even for more than two classes.

In the k-nearest-neighbor classifier method, a test sample is assigned to the class which is most frequently represented among the k nearest training samples. The nearest neighbors are determined by calculating the Euclidean distance function between those samples.

2.3.1. Support vector machine. SVM is a strong classification method which has demonstrated its excellent generalization property in various applications, e.g. the object recognition from an image (Osuna *et al* 1997), the classification of hand written characters (Oliveira and Sabourin 2004) and the speech recognition (Ganapathiraju *et al* 2004). Moreover, SVM have been also applied in BCI research by Müller *et al* (2003). To correctly predict the class to which an unseen test sample belongs, SVM calculates the decision hyperplane with the largest margin.

We may assume N training samples $x_i \in R^n$, each associated with a class label $y_i \in \{+1, -1\}$, $i = 1, \ldots, N$. The standard SVM solution is derived from the following optimizing problem (Cortes and Vapnik 1995):

$$\begin{gathered} \min_{w,b,\xi} \frac{1}{2} w^T w + c \sum_{i=1}^{N} \xi_i \\ \text{subject to } y_i (w^T \Phi(x_i) + b) \geqslant 1 - \xi_i, \\ \xi_i \geqslant 0, \qquad i = 1, \ldots, N, \end{gathered}$$

where w is the normal vector and b is the bias of the separation hyperplane. In case $\Phi(x) = x$ SVM is a linear classifier. Otherwise, if Φ maps x to a higher dimensional space, the SVM is termed nonlinear.

In case the training data cannot be separated without error, the slack variable $\xi_i \geqslant 0$ and the penalty parameter $C > 0$ have to be introduced. As a consequence, a training sample is allowed to be a small distance ξ_i on the wrong side of the hyperplane without violating the stated constraint. But the performance of SVM is not only determined by the available training samples, rather by the penalty parameter C as well. Therefore, the choice of an appropriate value of C is an essential part of the SVM classification method.

In practice, the optimization problem mentioned above is usually solved in its dual form, providing the advantage of simpler constraints:

$$\begin{gathered} \min_\alpha \tfrac{1}{2} \alpha^T Q \alpha - e^T \alpha \\ \text{subject to } 0 \leqslant \alpha_i \leqslant C, \qquad i = 1, \ldots, N, \\ y^T \alpha = 0, \end{gathered}$$

where e is the vector of all ones, Q is an $N \times N$ positive semidefinite matrix, $Q_{ij} = y_i y_j K(x_i, x_j)$, and $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$ is the kernel function. The decision function is

$$\mathrm{sgn}\left(\sum y_i \alpha_i K(x_i, x_j) + b \right).$$

2.3.2. Solving multi-class problems with binary classifiers. Since originally LDA and SVM were designed for binary classification problems, an extension for multiple classes is needed. A traditional and straightforward way to comply with this need is the combination of several binary classifiers to construct a multi-class classifier. Alternatively, algorithms can be used that consider all classes at once. For example, in the case of M-class LDA, as proposed by Duda *et al*

(2001), the n-dimensional feature vectors are projected into a $(n-1)$-dimensional subspace and a further classifier (e.g. a nearest-neighbor classifier) is used to create the full classifier. contrast, the combination of binary LDA does not need a second classifier. In the case of SVM, the optimization task of the combined SVM is smaller and easier to solve than a multi-class SVM. The combined SVM is almost as effective as the multi-class SVM, if the underlying binary classifiers are tuned appropriately (Rifkin and Klautau 2004). We have applied the combination of binary classifiers to our four-class EEG analysis for its simplicity.

Two approaches to extend binary classifiers are available to solve a M-class problem: the *one-versus-rest* and the *one-versus-one* schemes. In the one-versus-rest scheme M binary classifiers are constructed by training the ith classifier through labeling the samples of the ith class as positive and the remaining samples as negative. In the one-versus-one scheme, however, one classifier is constructed for each pair of classes. In total

$$\binom{k}{2} = \frac{k!}{(k-2)! \times 2!} = \frac{k \times (k-1)}{2}$$

binary classifiers are necessary for the different pairs of classes. For both schemes, the class of a test sample can be predicted by majority voting, i.e., the test sample is labeled by the class with the most numbers of votes. If ties or contradictions arise in the voting, the test sample can either be rejected or assigned to the class with largest prior probability. The latter was used in this analysis. Although the one-versus-one scheme seems slightly more complex than the one-versus-rest scheme, we preferred the one-versus-one scheme in our analysis, following the assumption that it may be more suitable for practical use (Hsu and Lin 2002).

3. Experiments and results

3.1. Evaluation criteria—the kappa coefficient

In an M-class classification problem, the proper evaluation of the classifier is described by its confusion matrix defining the relationship between the 'true' classes and the output of the classifier. From the confusion matrix H, we can derive the classification accuracy ACC (overall agreement)

$$\mathrm{ACC} = p_0 = \frac{\sum_i H_{ii}}{N}$$

and the chance expected agreement

$$p_e = \frac{\sum_i n_{oi} \times n_{io}}{N \times N},$$

where $N = \sum_i \sum_j H_{ij}$ is the total number of samples, H_{ii} are the elements of the confusion matrix H on the main diagonal, n_{oi}, n_{io} are the sums of each column and each row, respectively. Then the estimate of the kappa coefficient κ

$$\kappa = \frac{p_0 - p_e}{1 - p_e} = \frac{M \times p_0 - 1}{M - 1}$$

and its standard error $\mathrm{se}(\kappa)$ is obtained by

$$\mathrm{se}(\kappa) = \frac{\sqrt{p_o + p_e^2 - \sum_i [n_{oi} \times n_{io} \times (n_{oi} + n_{io})]/N^3}}{(1 - pe)\sqrt{N}}$$

with chance probability $p_e = 1/M$. For more details see also Cohen (1960), Bortz and Lienert (1998) and Kraemer (1982). To compute the kappa coefficient, we used the implementation realized in the BIOSIG-toolbox (Schlögl 2004).

3.2. Cross-validation

For cross-validation we chose a trial-based leave-one-out method (LOOM). We estimated the accuracy of a classifier by training the classifier m separate times, where m is the number of trials. Each time we removed one different trial from the previous data set and trained the classifier with the remaining trials. Then, we applied this developed classifier to each sample of the test trial. Accordingly, we calculated a classification result for each point in time and each trial, obtaining a time course of the classification result. LOOM was used in combination with each of the presented classifiers. All results presented here were obtained through this cross-validation procedure.

3.3. Single-channel analysis

In order to specify the importance of each channel (i.e., electrode position) for the classification result, the AAR parameters (model order $p = 3$) were estimated for every monopolar channel (total 60) and for every possible combination of bipolar channels ($60 \times 59/2 = 1770$). These bipolar channels were calculated by taking the difference of two monopolar channels. Accordingly, 1830 single-channel AAR estimates were obtained. Next, the AAR estimates (based on the data down sampled from 250 to 125 samples per second) from each trial were divided into segments of 25 samples (i.e. 0.2 s). For each segment an MDA classifier across all trials was calculated and applied to the same segment. Accordingly, an average kappa value for each segment was obtained. Within the interval of $t = 0$ to 7 s, the segment with the largest kappa value was used to obtain a classifier. The classifier was validated using leave-one-out method for cross-validation. This provides a time course of the kappa value for each of the 1830 channels. This algorithm is available through BIOSIG (Schlögl 2004).

The maximum kappa coefficient of each time course has been put into a 60×60 matrix at the position indicated by the respective channels (see first column of figure 3). The diagonal positions indicate the results from the monopolar channels, while all other positions display the results of bipolar channels. In order to validate the relative importance of each channel, the average of all maximum kappa coefficients over all channels was calculated. This average of all channels can be calculated by taking the average of all rows (or columns because of the symmetrical structure) of the 60×60 matrix. The average values are associated with each electrode and are displayed as a topographic map (second column of figure 3). The third column of figure 3 depicts the time course of the kappa coefficient for those three channels which yielded the largest kappa value, taking care that no channel was selected more than once. The legend lists those bipolar channels which provided the best results.

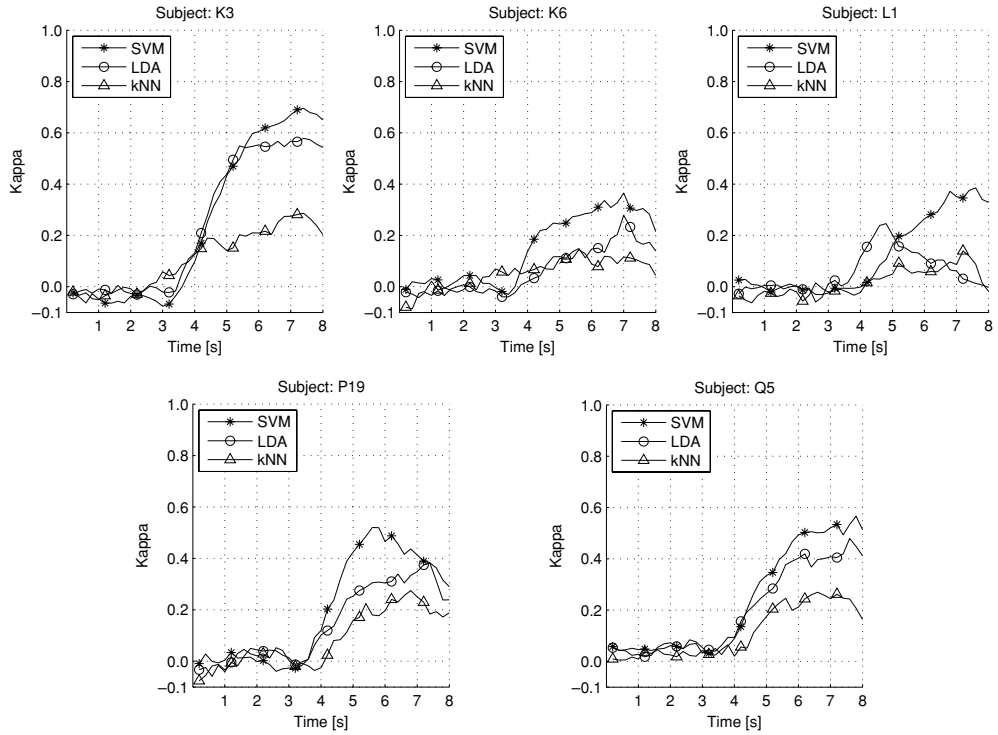

Figure 3. A comparison of kappa coefficient generated from linear SVM, LDA and *k*NN *output*. The classifiers were applied to analyze the four-class, 60-channel EEG recorded from subject K3, K6, L1, P19 and Q5 where the 180-dimensional feature vectors extracted by estimating the AAR-parameter of model order 3 were used as input.

Table 2. Classification results of the best single channel. TI indicates the segment for calculating the classifier, ACC indicates the overall accuracy, and $\kappa \pm \mathrm{se}(\kappa)$ indicates the kappa coefficient with its standard error. The positions of the channels are shown in figure 1.

Subject	Channels	TI (s)	ACC (%)	$\kappa \pm \mathrm{se}(\kappa)$
K3	#3–#34	6.6–6.8	56.9	0.425 ± 0.050
K6	#24–#41	5.0–5.2	46.5	0.288 ± 0.054
L1	#22–#33	6.2–6.4	48.5	0.313 ± 0.056
P19	#28–#17	5.8–6.0	54.4	0.392 ± 0.059
Q5	#30–#29	6.6–6.8	53.8	0.383 ± 0.051

The comparison of tables 2 and 3 reveals that in two data sets, K3 and L1, the three-channel case provide better results than the single-channel case, but with K6 the single-channel result is more advantageous. In both tables, K6 is indicated to show the worst overall performance and no prominent electrode over the sensory-motor area was determined (figure 3). In general, all data sets with high performance levels reach even higher results by adding more channels. Adding more channels to the data sets with low performance, however, did not increase the classification result.

3.4. Multi-channel analysis

The multi-channel analysis is based on all recorded 60 channels, for each of them the features were determined by estimating the AAR parameters (model order $p = 3$, update coefficient UC $= 0.0055$ and AAR parameters estimated by Kalman filtering). The EEG of each trial was represented as a $n \times S$ matrix, where $n = 60 \times 3 = 180$ constitutes the dimension of the samples. After the down-sampling from 250 Hz to 125 Hz each 8 s trial gives rise to 125 Hz $\times$ 8 s $=$ 1000 samples. As in the single-channel analysis, each trial was divided into smaller time segments with a length of 0.2 s (25 samples) and the separability for each segment was calculated.

Table 3. Classification result using the AAR parameters of the three best single channels; a model order of 3 was used for each channel resulting in nine features. The 'channel' column shows the three best single channels for each subject; TI indicates the segment for calculating the classifier, ACC indicates the overall accuracy, and $\kappa \pm \mathrm{se}(\kappa)$ indicates the kappa coefficient with its standard error. The channel positions are shown in figure 1.

Subject	Channels	TI (s)	ACC (%)	$\kappa \pm \mathrm{se}(\kappa)$
K3	#34–#3, #23–#14, #19–#12	5.8–6.0	66.6	0.555 ± 0.054
K6	#41–#24, #50–#34, #49–#25	6.8–7.0	38.5	0.180 ± 0.049
L1	#33–#22, #44–#31, #34–#28	6.4–6.6	49.5	0.327 ± 0.056
P19	#18–#17, #40–#39, #42–#34	6.0–6.2	57.7	0.432 ± 0.062
Q5	#30–#29, #24–#16, #23–#22	6.4–6.4	55.2	0.402 ± 0.052

Table 4. Best classification results using LDA, *k*NN and linear SVM. TI indicates the segment (time point) for calculating the classifier, ACC indicates the overall accuracy, and $\kappa \pm \mathrm{se}(\kappa)$ indicates the kappa coefficient with its standard error. The number of the nearest neighbor k used in *k*NN classifier and the penalty parameter C used in SVM are given in brackets in the first column of *k*NN and SVM part respectively. These results show that SVM is more suitable to classify high-dimensional data sets.

	LDA			*k*NN			SVM		
Subject	TI (s)	ACC (%)	$\kappa \pm \mathrm{se}(\kappa)$	TI (s) (k)	ACC (%)	$\kappa \pm \mathrm{se}(\kappa)$	TI (s) (C)	ACC (%)	$\kappa \pm \mathrm{se}(\kappa)$
K3	5.6–5.8	68.4	0.578 ± 0.055	6.8–7.0 (50)	46.5	0.287 ± 0.045	6.6–6.8 (1)	77.2	0.695 ± 0.059
K6	6.0–6.2	45.9	0.278 ± 0.054	5.0–5.2 (50)	35.9	0.145 ± 0.047	5.6–5.8 (10)	52.4	0.366 ± 0.058
L1	4.4–4.6	43.5	0.246 ± 0.052	6.8–7.0 (50)	35.5	0.140 ± 0.046	6.6–6.8 (100)	53.9	0.386 ± 0.059
P19	6.6–6.8	53.6	0.381 ± 0.059	6.8–7.0 (100)	45.6	0.275 ± 0.054	5.0–5.2 (100)	64.0	0.520 ± 0.065
Q5	6.2–6.4	60.9	0.479 ± 0.055	5.8–6.0 (100)	45.2	0.270 ± 0.046	6.0–6.2 (50)	67.5	0.566 ± 0.058

Table 5. Statistical significant differences (t-value) between classifiers. A positive t-value indicates that the column classifier is better than the classifier indicated in the corresponding row; a negative t-value indicates that classifier of the corresponding column is better than the classifier of the row.

t-Value	MDA (1)	MDA (3)	LDA (60)	*k*NN (60)	SVM (60)
MDA (1)	NA	0.50	0.80	−12.73 (***)	3.98 (**)
MDA (3)	−0.50	NA	0.38	−4.11 (**)	5.39 (**)
LDA (60)	−0.80	−0.37	NA	−4.71 (**)	9.80 (***)
*k*NN (60)	12.73 (***)	4.11 (**)	4.71 (**)	NA	8.45 (***)
SVM (60)	−3.98 (**)	−5.39 (**)	−9.80 (***)	−8.45 (***)	NA

One, two and three stars indicate that the significance level has reached $p = 5\%$, 1% and 0.1%, respectively.

We combined the binary LDA and SVM classifiers by using the one-versus-one scheme and the majority voting. Whenever ties during the voting process emerged, we allocated the sample to the class which had reached the maximal sum of all decision values (i.e., the distance between the decision hyper plane and the sample point). The classification results are summarized in a sequence of confusion matrices; specifically we extracted the time course of accuracy or the one of the kappa coefficient.

For *k*NN, the following values were chosen as numbers for the nearest neighbor k: 5, 10, 50, 100 and 200. The SVM penalty parameter C was set to the values 1, 10, 50 and 100. Table 4 shows the comparison of the most accurate LDA, *k*NN and linear SVM classification results, which were calculated individually for each subject. The corresponding parameters of the results are listed in the first column for each classifier.

Figure 3 represents the classification results of the four-classe EEG as time courses of kappa coefficients. These results indicate SVM as more effective than LDA that was in turn more efficient than *k*NN. In most cases, the time course of the kappa coefficients reveals a similar characteristic differing only in the overall performance. An exception is the result of subject L1 insofar as the peak of the LDA-based time course appears much earlier ($t = 4.5$ s), whereas the SVM-based peak does much later ($t = 6.7$ s).

3.5. Statistical comparison

In total, results from five different classifiers have been obtained. First, AAR parameters ($p = 3$) of each possible bipolar channel were applied to MDA (table 2). Second, the AAR parameters of the three channels with the best classification results were also applied to MDA (table 3). Third, the AAR parameters obtained from each of the 60 channels were applied to LDA, k-nearest neighbor and an SVM classifier. The kappa values together with their confidence intervals (see tables 2, 3 and 4) prove that all classification results are significantly above a chance classification.

In order to compare the different classification approaches the accuracy and kappa values (listed in tables 2, 3 and 4) were subjected to statistical analysis. The t-value of the differences between each pair of classifiers were calculated; with 4 degrees of freedom a t-value larger than 2.78 is with error probability $p < 5\%$ statistically significant.

Table 5 contains the t-values resulting from the significance test of the differences between classifiers. Accordingly, SVM reached significantly better results than any of the four other classifiers, whereas *k*NN was significantly worse than the rest. The difference between single-channel MDA, three-channel MDA, and LDA with all 60 channels failed to reach statistical significance. Table 5 lists the results of the kappa coefficients. The same analysis performed with the ACC criterion yielded the same results.

4. Discussion and conclusion

Five different classification approaches were applied and compared with each other: single-channel MDA, MDA based on the three best channels, LDA using 60 channels, SVM using 60 channels and *k*NN using 60 channels. Looking at both linear classifiers, SVM and LDA, SVM performed significantly better than LDA. Two related explanations should be raised:

(i) The three AAR parameters were calculated in this study for each of the 60 channels resulting in a total feature space of 180 dimensions, while only 240 to 320 trials were used to build a classifier. The ratio between 180 dimensions and

less than twice as many independent samples demonstrate that the 'curse of dimensionality' cannot be neglected. To overcome the 'curse of dimensionality' problem is one of the main advantages of SVM.

(ii) LDA requires the estimation of the inverse of the covariance matrix to determine its weight factors. If the dimensionality of the data is high and only few samples are available, then, generally, the estimate of the covariance matrix and its inverse is bad (Duda *et al* 2001). SVM does not depend on the covariance matrix and, therefore, is not affected by this limitation.

LDA and MDA performed significantly worse than SVM. However, the advantage of these two statistical classifiers is their low computational effort. We used this advantage to investigate all possible bipolar and monopolar channels. In the face of the similarity of the MDA and LDA results no statistically significant difference is observed. This observation complies with the general property of robustness of statistical classifiers.

As indicated by our results, the neural network *k*NN performed worst, even worse than the single-channel result. A previous study pointed to the same direction (Schlögl 2000, pp 36–7) showing that learning vector quantization (LVQ) reached a lower degree of accuracy than LDA. Given that both, LVQ and *k*NN, are the classifiers based on neural networks. This raises the question, whether neural network-based classifiers are in general less suited to classify AAR parameters.

Often it is more convenient for the subject to use only a limited number of electrodes (i.e. channels). Therefore, to evaluate each single channel in its discriminative power we performed a minimum distance analysis based on the mahalanobis distance (MDA). The results revealed a similar performance as compared to LDA with multi-channel data.

The single-channel analysis uncovers the relative importance of each electrode position. The dominance of electrodes in the single trial analyses (e.g. figure 3) that overlay the sensorimotor and premotor areas and especially the hand representation area confirms the modulation of sensorimotor rhythms during motor imagery. McFarland *et al* (1997) used the r^2 to measure the proportion of the total variance of the mu and or beta rhythm amplitude accounting for the user's target location and underlining the importance of electrode locations at or close to C3 and C4 for BCI applications. In another Q1 single-trial motor imagery study Pregenzer *et al* (1995) applied the distinction sensitive learning vector quantizer (DSLVQ) and reported the best separability between left and right hand motor imagery with signals recorded from electrode positions C3 and C4. The activation of sensorimotor areas during movement imagination could be quantified by EEG studies (e.g., Neuper and Pfurtscheller 1999) but was measured also indirectly in fMRI studies (e.g., Ball *et al* 1999, Porro *et al* 1996). All this evidence underlines the importance of the Rolandic mu and central beta rhythms for the realization of an EEG-based BCI and the attainment of control over brain oscillations.

The current work made use of the advantageous method of AAR parameters, allowed a low-dimensional feature space (compared to other spectral features), and no feature selection was necessary. The model order $p = 3$ and the update coefficient were selected according to unpublished results from different data sets using an MDA classifier. This study did not include any optimization of the model order nor of the update coefficient. Therefore, as a matter of fact, optimization of the model order and/or the update coefficient provides room for further improvement.

Given the fact that a rather limited number of independent trials (60 to 90) were available for each class, we implemented a cross-validation using a trial-based leave-one-out method (LOOM). LOOM was applied successfully already by one of the earliest works in the field of BCI research (Keirn and Aunon 1990). The striking advantage of LOOM rests on its ability that even in the presence of small sample sizes efficient and unbiased cross-validation is reached.

Generally, the usefulness or rather the property of a classifier with several classes is characterized by the confusion matrix. From the confusion matrix the accuracy as well as Cohen's kappa coefficient (Cohen 1960) can be derived. If all M classes occur equally frequent, kappa κ and accuracy ACC can be related to this equation:

$$\kappa = \frac{M \times \mathrm{ACC} - 1}{M - 1}.$$

However, if their occurrence is not equally distributed, the kappa coefficient has to be preferred since in this way the higher rates of occurrence of some classes are compensated. In other areas such as in sleep classification research (e.g. Danker-Hopfe *et al* 2004, Anderer *et al* 2005) the kappa coefficient is already established. The current work is an attempt to introduce Cohen's kappa to the field of BCI research.

In summary, this four-class BCI data of five subjects were examined on various aspects: determining the overall separability with statistical, neural network-based and support-vector-machine-based classifiers, comparing these approaches with each other and, finally to estimate the discriminative power of specific electrode positions.

Three data sets (K3, K6 and L1) have also been used for the BCI competition 2005. It will be of interest to compare the results of this work with the submissions of the BCI competition. Overall, a classification accuracy of between 52% and 77% and a kappa between 0.36 and 0.70 were obtained.

Acknowledgments

The work was partly funded by the Austrian grant 'Fonds zur Förderung der wissenschaftlichen Forschung', project number P16326-B02. The authors thank Claudia Keinrath for data recording and Gernot Supp for proofreading.

References

Anderer P *et al* 2005 An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24 × 7 utilizing the Siesta Database *Neuropsychobiology* **51** 115–33

Ball T, Schreiber A, Feige B, Wagner M, Lucking C H and Kristeva-Feige R 1999 The role of higher-order motor areas in voluntary movement as revealed by high-resolution EEG and fMRI *Neuroimage* **10** 682–94

Birbaumer N, Hinterberger T, Kubler A and Neumann N 2003 The thought-translation device (TTD): neurobehavioral mechanisms and clinical outcome *IEEE Trans. Neural Syst. Rehabil. Eng.* **11** 120–3

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

Bortz J and Lienert G A 1998 *Kurzgefasste Statistik für die klassische Forschung (Kapitel 6: Uebereinstimmungsmasze fuer subjektive Merkmalsurteile)* (Berlin: Springer) pp 265–70

Q2 Chang C-C and Lin C-J 2001 LIB-SVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm

Cortes C and Vapnik V 1995 Support-vector networks *Mach. Learn.* **20** 273–97

Cohen J 1960 A coefficient of agreement for nominal scales *Educ. Psychol. Meas.* **20** 37–46

Danker-Hopfe H *et al* 2004 Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders *J. Sleep Res.* **3** 63–9

Duda R O, Hart P E and Stort D G 2001 *Pattern Classification* (New York: Wiley)

Ganapathiraju A, Hamaker J E and Picone J 2004 Applications of support vector machines to speech recognition *IEEE Trans. Signal Process.* (see also *IEEE Trans. Acoust., Speech Signal Process.* **52** 2348–55)

Hsu C W and Lin C J 2002 A comparison of methods for multiclass support vector machines *IEEE Trans. Neural Netw.* **13** 415–25

Keirn Z A and Aunon J I 1990 A new mode of communication between man and his surroundings *IEEE Trans. Biomed. Eng.* **37** 1209–14

Kraemer H C 1982 Kappa coefficient *Encyclopedia of Statistical Sciences* ed S Kotz and N L Johnson (New York: Wiley)

McFarland D J, McCane L M, David S V and Wolpaw J R 1997 Spatial filter selection for EEG-based communication *Electroencephalogr. Clin. Neurophysiol.* **103** 386–94

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

Neuper C and Pfurtscheller G 1999 Motor imagery and ERD *Event-Related Desynchronization—Handbook of Electroencephalography and Clinical Neurophysiology* vol 6 revised edn, ed G Pfurtscheller and F H Lopes da Silva (Amsterdam: Elsevier)

Oliveira L S and Sabourin R 2004 Support vector machines for handwritten numerical string recognition *Proc. 9th Int. Workshop on Frontiers in Handwriting Recognition* vol 9, pp 39–44

Osuna E, Freund R and Girosit F 1997 Training support vector machines: an application to face detection *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* pp 130–6

Pfurtscheller G *et al* 2003 Graz-BCI: state of the art and clinical applications *IEEE Trans. Neural Syst. Rehabil. Eng.* **11** 177–80

Pfurtscheller G, Flotzinger D, Pregenzer M, Wolpaw J R and McFarland D 1995 EEG-based brain computer interface (BCI): search for optimal electrode positions and frequency components *Med. Prog. Technol.* **21** 111–21

Porro C A, Francescato M P, Cettolo V, Diamond M E, Baraldi P, Zuiani C, Bazzocchi M and di Prampero P E 1996 Primary motor and sensory cortex activation during motor performance and motor imagery: a functional magnetic resonance imaging study *J. Neurosci.* **16** 7688–98

Rifkin R and Klautau A 2004 In defense of one-vs-all classification *J. Mach. Learn. Res.* **5** 101–41

Schlögl A, Lugger K and Pfurtscheller G 1997 Using adaptive autoregressive parameters for a brain–computer-interface experiment *Proc. 19th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* vol 4 pp 1533–5

Schlögl A 2000 *The Electroencephalogram and the Adaptive Autoregressive Model: Theory and Applications* (Aachen, Germany: Shaker Verlag)

Schlögl A 2004 BIOSIG—an Open Source Software Library for biomedical Signal Processing. Available online at http://biosig.sourceforge.net

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

Queries

(1) Author: Please include 'Pregenzer *et al* (1995)' in the reference list.
(2) Author: Please cite 'Chang and Lin (2001)' and 'Pfurtscheller *et al* (1995) in text.