Support Vector Channel Selection in BCI

Thomas Navin Lal*, *Student Member, IEEE*, Michael Schröder, Thilo Hinterberger, Jason Weston, Martin Bogdan, Niels Birbaumer, and Bernhard Schölkopf

***Abstract*—Designing a brain computer interface (BCI) system one can choose from a variety of features that may be useful for classifying brain activity during a mental task. For the special case of classifying electroencephalogram (EEG) signals we propose the usage of the state of the art feature selection algorithms Recursive Feature Elimination [3] and Zero-Norm Optimization [13] which are based on the training of support vector machines (SVM) [11]. These algorithms can provide more accurate solutions than standard filter methods for feature selection [14].**

We adapt the methods for the purpose of selecting EEG channels. For a motor imagery paradigm we show that the number of used channels can be reduced significantly without increasing the classification error. The resulting best channels agree well with the expected underlying cortical activity patterns during the mental tasks.

Furthermore we show how time dependent task specific information can be visualized.

***Index Terms*—Brain computer interface (BCI), channel relevance, channel selection, electroencephalography (EEG), feature relevance, feature selection, Recursive Feature Elimination (RFE), support vector machine (SVM), Zero Norm Optimization (l0-Opt).**

I. Introduction

MOST brain computer interfaces (BCIs) make use of mental tasks that lead to distinguishable electroencephalogram (EEG) signals of two or more classes. For some tasks the relevant recording positions are known, especially when the tasks comprise motor imagery, e.g., the imagination of limb movements, or the overall activity of large parts of the cortex that occurs during intentions or states of preparation and relaxation.

For the development of new paradigms whose neural correlates are not known in such detail, finding optimal recording positions for use in BCIs is challenging. New paradigms can become necessary when motor cortex areas show lesions, for the increase of the information rate of BCI systems or for robust multi-class BCIs. If good recording positions are not known, a simple approach is to use data from as many as possible EEG electrodes for signal classification. The drawback of this approach is that the extend to which feature selection and classification algorithms overfit to noise increases with the number of task-irrelevant features, especially when the ratio of training points and number of features is small. In addition, it is difficult to understand which part of the brain generates the class relevant activity.

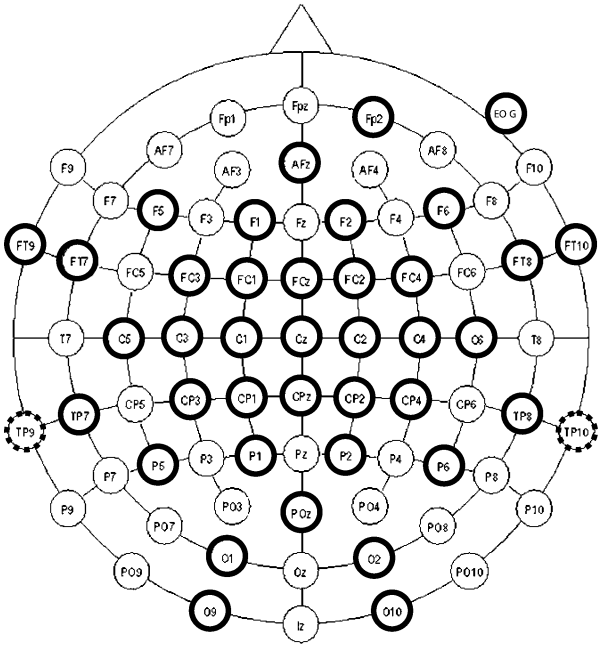


Fig. 1. The position of 39 EEG electrodes used for data acquisition are marked in solid black circles. The two referencing electrodes are marked in dotted circles.

We show that the selection of recording positions can be done robustly in the absence of prior knowledge about the spatial distribution of brain activity of a mental task. Specifically we adapt the state of the art feature selection methods *Zero-Norm Optimization* (l0-Opt) and *Recursive Feature Elimination* (RFE) to the problem of channel selection and demonstrate the usefulness of these methods on the well known paradigm of motor imagery.

The paper is structured as follows: Section II contains the experimental setup, the task, and the basic data preprocessing. In Section III, the feature selection methods and the classification algorithm are described. Results are given in Section IV and the final section concludes.

II. Data Acquisition

A. Experimental Setup and Mental Task

We recorded EEG signals from eight untrained right handed male subjects using 39 silver chloride electrodes (see Fig. 1). The reference electrodes were positioned at TP9 and TP10. The

Manuscript received July 16, 2003; revised March 17, 2004. This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG) and in part by the National Institute of Health (NIH). *Asterisk indicates corresponding author*.

*T. N. Lal is with Max-Planck-Institut for Biological Cybernetics, Spemannstr. 38, Tübingen 72076, Germany (e-mail: navin@tuebingen.mpg.de).

M. Schröder and M. Bogdan are with Eberhard Karls University Tübingen, Department of Computer Engineering, Tübingen 72076, Germany (e-mail: schroedm@informatik.uni-tuebingen.de; bogdan@informatik.uni-tuebingen.de).

T. Hinterberger and N. Birbaumer are with Eberhard Karls University Tübingen, Institute of Medical Psychology and Behavioral Neurobiology, Tübingen 72076, Germany (e-mail: thilo.hinterberger@uni-tuebingen.de; niels.birbaumer@uni-tuebingen.de).

J. Weston and B. Schölkopf are with Max-Planck-Institut for Biological Cybernetics, Tübingen 72076, Germany (e-mail: jason.weston@tuebingen.mpg.de; bs@tuebingen.mpg.de).

Digital Object Identifier 10.1109/TBME.2004.827827

two electrodes Fp2 and 1 cm lateral of the right eye (EOG) were used to record possible EOG artifacts and eye blinks while two fronto-temporal and two occipital electrodes were positioned to detect possible muscle activity during the experiment. Before sampling the data at 256 Hz an analog bandpass filter with cutoff frequencies 0.1 Hz and 40 Hz was applied.

The subjects were seated in an armchair at 1-m distance in front of a computer screen. Following the experimental setup of [6] the subjects were asked to imagine left versus right hand movements during each trial. With every subject, we recorded 400 trials during one single session. The total length of each trial was 9 s. Additional intertrial intervals for relaxation varied randomly between 2 and 4 s. No outlier detection was performed and no trials were removed during the data processing at any stage.

Each trial started with a blank screen. A small fixation cross was displayed in the center of the screen from second 2 to 9. A cue in the form of a small arrow pointing to the right or left side was visible for half a second starting with second 3. In order to avoid event related signals in later processing stages, only data from seconds 4 to 9 of each trial was considered for further analysis. Feedback was not provided at any time.

B. Preanalysis

As Pfurtscheller and da Silva have reported [7] that movement related desynchronization of the μ-rhythm (8–12 Hz) is not equally strong in subjects and might even fail for various reasons (e.g., because of too short intertrial intervals that prevent a proper re-synchronization) we performed a preanalysis in order to identify and exclude subjects that did not show significant μ-activity at all.

For seven of the eight subjects the μ-band was only slightly differing from the 8–12 Hz usually given in the EEG literature. Only one subject showed scarcely any activity in this frequency range but instead a recognizable movement related desynchronization in the 16–20 Hz band.

Restricted to only the 17 EEG channels that were located over or close to the motor cortex we calculated the maximum energy of the μ-band using the Welch method [12] for each subject. This feature extraction resulted in one parameter per trial and channel and explicitly incorporated prior knowledge about the task.

The eight data sets consisting of the Welch-features were classified with linear SVMs (see below) including individual model selection for each subject. Generalization errors were estimated by tenfold cross validation (CV). As for three subjects the preanalysis showed very poor error rates close to chance level their data sets were excluded from further analysis.

C. Data Preprocessing

For the remaining five subjects the recorded 5 s windows of each trial resulted in a time series of 1280 sample points per channel. We fitted an autoregressive (AR) model of order 3 to the time series[1] of all 39 channels using forward backward linear prediction [5]. The three resulting coefficients per channel and trial formed the new representation of the data.

The extraction of the features did not explicitly incorporate prior knowledge although AR models have successfully been used for motor related tasks (e.g., [6]). However, they are not directly linked to the μ-rhythm.

D. Notation

Let n denote the number of training vectors (trials) of the data sets ($n = 400$ for all five data sets) and let d denote the data dimension ($d = 3 \cdot 39 = 117$ for all five data sets). The training data for a classifier is denoted as $X = (x^{(1)}, \ldots, x^{(n)}) \in \mathbb{R}^{n \times d}$ with labels $Y = (y_1, \ldots, y_n) \in \{-1, 1\}^n$. For the task used in this paper $y = -1$ denotes imagined left hand movement, $y = 1$ denotes imagined right hand movement. The terms *dimension* and *feature* are used synonymously. For $l \in \mathbb{N}$, $l > 1$ the set $M^{-j} \subset \mathbb{R}^{l-1}$ is obtained from a set $M \subset \mathbb{R}^l$ by removing the dimension j from every point $m \in M$ (canonical projection).

III. Feature Selection and Classification Methods

Feature selection algorithms can be characterized as either filter or wrapper methods [8]. They select or omit dimensions of the data depending on a performance measure.

The problem of how to rate the relevance of a feature if nonlinear interactions between features are present is not trivial, especially since the overall accuracy might not be monotonic in the number of features used. Some feature selection methods try to overcome this problem by optimizing the feature selection for subgroups of fixed sizes (plus-l take-away-r search) or by implementing floating strategies (e.g., floating forward search) [8]. Only few algorithms like, e.g., genetic algorithms can choose subgroups of arbitrary size during the feature selection process. They have successfully been used for the selection of spatial features [10] in BCI applications but are computationally demanding.

For the application of EEG channel selection, it is necessary to treat a certain kind of grouped features homogenously: numerical values belonging to one and the same EEG channel have to be dealt with in a congeneric way so that a spatial interpretation of the solution becomes possible. We adapted the state of the art feature selection methods 10-Opt and RFE as well as the Fisher Correlation to implement these specific requirements. The first two algorithms are closely related to SVMs.

A. Support Vector Machines (SVMs)

The SVM is a relatively new classification technique developed by Vapnik [11] which has shown to perform strongly in a number of real-world problems, including BCI [2]. The central idea is to separate data $X \subset \mathbb{R}^d$ from two classes by finding a weight vector $w \in \mathbb{R}^d$ and an offset $b \in \mathbb{R}$ of a hyperplane

$$H : \mathbb{R}^d \to \{-1, 1\}$$
$$x \mapsto \operatorname{sign}(w \cdot x + b)$$

with the largest possible margin,[2] which apart from being an intuitive idea has been shown to provide theoretical guaranties in

[1]For this choice, we compared different model orders. For a given order, we fitted an AR-model to each EEG sequence. After proper model selection a support vector machine (SVM) with tenfold CV was trained on the coefficients. Model order 3 resulted in the best mean CV error.

[2]Is X linear separable the margin of a hyperplane is the distance of the hyperplane to the closest point $x \in X$.

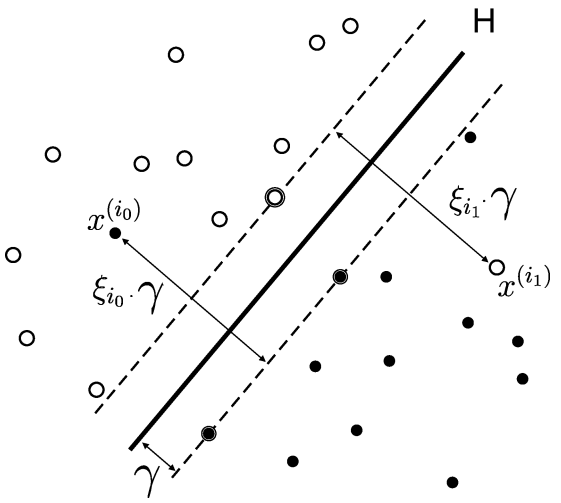


Fig. 2. Linear SVM. For nonseparable data sets, slack variables ξ_i are introduced. The thick points on the dashed lines are called support vectors (SVs). The solution for the hyperplane H can be written in terms of the SVs. For more detail see Section III-A.

terms of generalization ability [11]. One variant of the algorithm consists of solving the following optimization problem:

$$\begin{aligned} \min_{w\in\mathbb{R}^d} \quad & \|w\|_2 + C\sum_{i=1}^{n}\xi_i^2 \\ \text{s.t.} \quad & y_i(w\cdot x^{(i)}+b) \geq 1-\xi_i \quad (i=1,\ldots,n) \end{aligned} \tag{1}$$

The parameters ξ_i are called slack variables and ensure that the problem has a solution in case the data is not linear separable[3] (see Fig. 2). The margin is defined as $\gamma(X,Y,C) = 1/\|w\|_2$. In practice, one has to trade-off between a low training error, e.g., $\sum\xi_i^2$, and a large margin γ. This trade-off is controlled by the regularization parameter C. Finding a good value for C is part of the model selection procedure. If no prior knowledge is available C has to be estimated from the training data, e.g., by using CV. The value $2/C$ is also referred to as the *ridge*. For a detailed discussion please refer to [9].

B. Fisher Criterion (FC)

The FC determines how strongly a feature is correlated with the labels [1]. For a set $T = \{t^{(1)},\ldots,t^{(|T|)}\} \subset \mathbb{R}^d$ define the mean $\mu_j(T) = (1)/(|T|)\sum_{i=1}^{|T|} t_j^{(i)}$ and the variance $V_j(T) = (1)/(|T|)\sum_{i=1}^{|T|}(t_j^{(i)}-\mu_j(T))^2$ $(j=1,\ldots,d)$. The score R_j of feature j is then given by

$$R_j(X) = \frac{(\mu_j(X^+)-\mu_j(X^-))^2}{V_j(X^+)+V_j(X^-)} \tag{2}$$

with $X^+ := \{x_i \in X | y_i = 1\}$ and X^- similarly. The rank of a channel is simply set to the mean score of the corresponding features.

C. Zero-Norm Optimization (l0-Opt)

Weston *et al.* [13] recently suggested to minimize the zero-norm[4] $\|w\|_0 := \text{cardinality}(\{w_j : w_j \neq 0\})$ instead of minimizing the l_1-norm or l_2-norm as in standard SVMs [see, for example (1)]

$$\begin{aligned} \min_{w\in\mathbb{R}^d} \quad & \|w\|_0 + C\|\xi\|_0 \\ \text{s.t.} \quad & y_i\left(w\cdot x^{(i)}+b\right) \geq 1-\xi_i \,(i=1,\ldots,n). \end{aligned} \tag{3}$$

The solution of this optimization problem is usually much sparser than the solution of problem (1). Thus, feature selection is done implicitly. Unfortunately the problem has shown to be NP-hard but the authors developed an iterative method to approximate the solution. In case the solution w^* has less than the desired number of zero entries, the remaining features $\{j\}$ can be ranked according to w_j^* (as in one iteration step of RFE).

In the original version of the method, the features are multiplied with a scaling factor during each iteration. Once a scaling factor is zero, the corresponding feature is removed. We adapt this method in the following way: the scaling factors of the features corresponding to a channel are substituted by their mean. Thus, all features of one channel are either removed completely (the channel is removed) or all features remain. As in the case of SVM and RFE, the parameter C has to be estimated from the training data in case prior knowledge is not available.

D. Recursive Feature Elimination (RFE)

This feature selection method was prosed by Guyon *et al.* [4] and is based on the concept of margin maximization. The importance of a dimension is determined by the influence it has on the margin of a trained SVM. Let W be the inverse of the margin

$$W(X,Y,C) := \frac{1}{\gamma(X,Y,C)} = \|w\|_2.$$

At each iteration one SVM is trained and the features $\hat{j}$ which minimize $|W(X,Y,C) - W(X^{-j},Y^{-j},C)|$ are removed (typically that is one feature only); this is equivalent to removing the dimensions $\hat{j}$ that correspond to the smallest $|w_j|$. We adapt this method for channel selection in the following way: Let $F_k \subset \{1,\ldots,d\}$ denote the features from channel k. Similar to the reformulation of the FC and the 10-Opt, we define for each channel k the score $s_k := 1/|F_k|\sum_{l\in F_k}|w_l|$. At each iteration step we remove the channel with the lowest score. The parameter C has to be estimated from the training data, if no prior knowledge is available.

For the remainder of the paper we refer to the adapted feature selection methods as channel selection methods. Furthermore, we will also refer to the adapted RFE as *Recursive Channel Elimination*.

E. Generalization Error Estimation

For model selection purposes we estimated the generalization error of classifiers via tenfold CV.

If the generalization error of a channel selection method had to be estimated, a somewhat more elaborated procedure was used. An illustration of this procedure is given in Fig. 3.

[3]Is the data linear separable the slack variables can improve the generalization ability of the solutions.

[4]The zero-norm of a vector v is equal to number of nonzero entries of v.

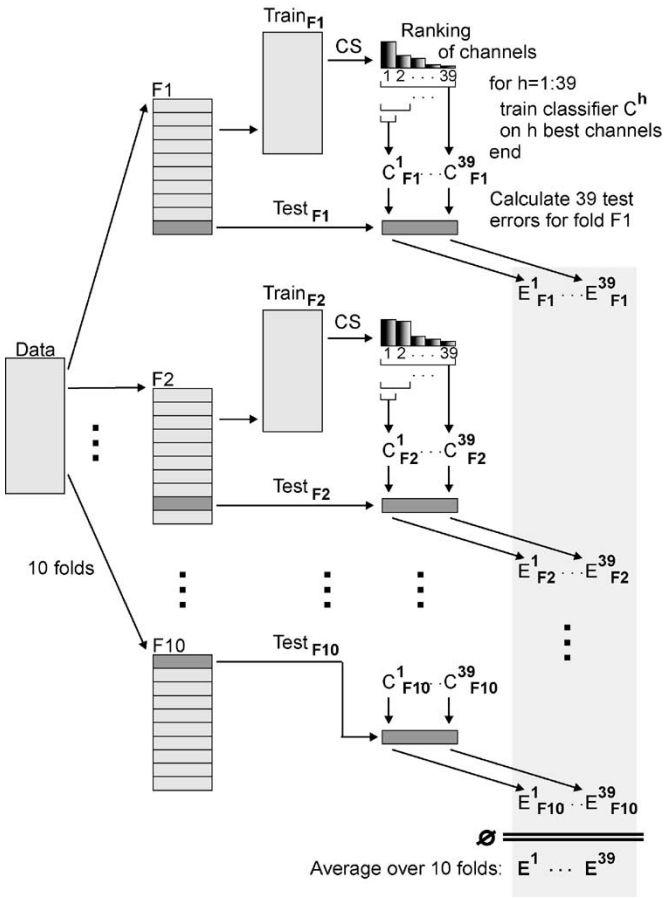


Fig. 3. Illustration of the procedure for channel selection and error estimation using CV.

The whole data set is split up into 10 folds (*F1* to *F10*) as for usual CV. In each fold F, the channel selection (CS in Fig. 3) is performed based on the train set of F only, leading to a specific ranking of the 39 EEG channels. For each fold F, 39 classifiers C_F^h, $h = 1, \ldots, 39$ are trained as follows: C_F^h is trained on the h best[5] channels, respectively, of the train set of F and tested on the corresponding channels of the test set of F. For each fold, this results in 39 test errors (E_F^1 to E_F^{39}).

During the last step, the corresponding test errors are averaged over all folds. This leads to an estimate of the generalization error for every number of selected channels.

IV. Results

A. Channel Selection

We applied the three channel selection methods FC, RFE, and 10-Opt introduced in Section III to the five data sets. As the experimental paradigm is well known, we could examine the results concerning their physiological plausibility. Therefore, we investigated whether the best ranked channels are those situated over or close to motor areas. Furthermore we analyzed if the number of channels can be reduced without a loss of accuracy in terms of CV error.

[5]In this context, *best* means according to the calculated ranking of that fold.

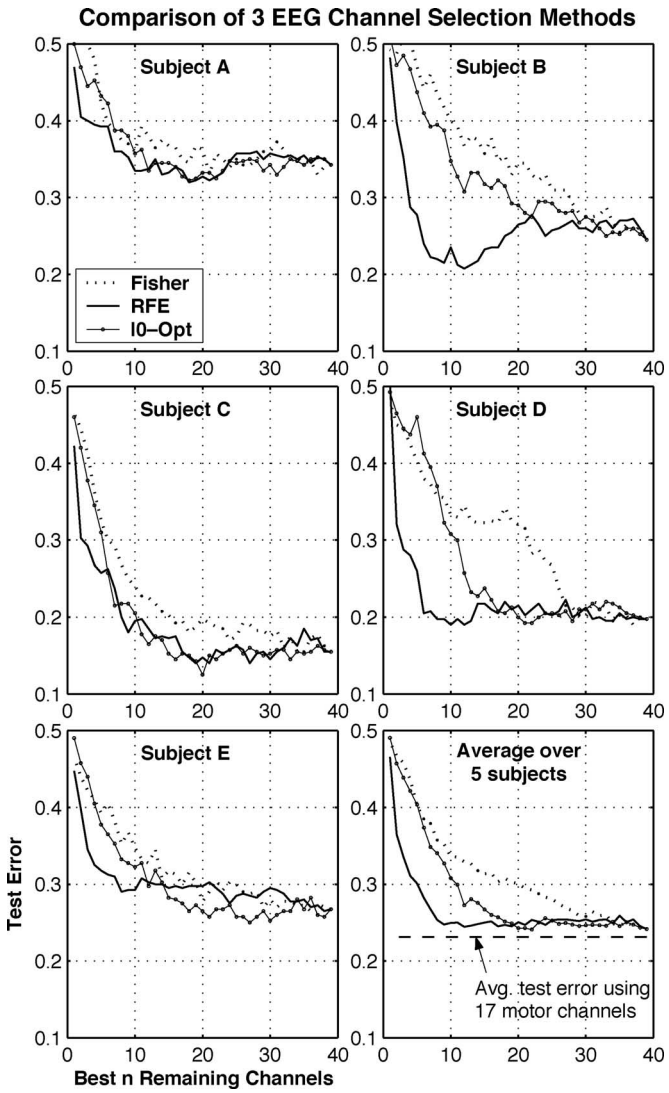


Fig. 4. Comparison of the three channel selection methods *Fisher Score*, *RFE* and *l0-Opt* individually for five subjects and averaged over the subjects. Method RFE allows the strongest reduction of number of channels for all subjects.

Initial to the channel selection and individually for each subject s, the regularization parameter C_s for later SVM trainings was estimated via tenfold CV from the training data sets.[6]

The estimation of the generalization error for all 39 stages of the channel selection process[7] was carried out using linear SVMs as classifiers with parameters C_s previously determined. Details about the tenfold CV during the estimation procedure are described in Section III-E and Fig. 3.

The estimation results are depicted in Fig. 4. The first five plots show the individual generalization error for the five subjects against the different numbers of channels chosen by the three channel selection methods. The sixth plot in the bottom right corner shows the generalization error of the three methods averaged over the five subjects.

[6]Estimating the parameter for each number of channels in the process of channel selection might improve the accuracy. However the chance of overfitting increases.

[7]In fact, methods RFE and l0-Opt perform rather a channel *removal* than a channel selection.

TABLE I
RFE RANKING OF 39 EEG CHANNELS

	Subjects				
Rank	A	B	C	D	E
1	C4	CP4	CP4	FC4	CP4
2	CP4	C3	CP3	C4	CPz
3	CP2	C4	C4	CP2	C2
4	C2	FC4	C2	CP1	FC3
5	Cz	FT9	C1	C3	C4
6	FC4	FT10	CPz	FC3	C1
7	FC2	CP1	CP2	C2	FCz
8	C3	C1	C3	C1	FC4
9	CP3	F6	F1	FC2	C3
10	F1	Fp2	FC1	FC1	POz
11	F2	FC1	FC2	FT10	P6
12	C1	AFz	C5	FCz	O10
13	FC3	C2	FT7	F2	FC1
14	CPz	P6	F2	FT9	C6
15	CP1	CP2	FC3	F1	C5
16	FCz	P1	C6	C5	Cz
17	P2	EOG	P1	F5	CP2
18	P1	FC3	CP1	C6	O1
19	C6	Cz	O1	POz	O9
20	AFz	C6	POz	AFz	TP8
21	F5	TP8	TP7	FT8	CP1
22	C5	P2	Fp2	Fp2	P1
23	FT9	POz	P5	P2	F1
24	FC1	F2	P6	P1	F2
25	FT7	FC2	FC4	O10	FT7
26	POz	O10	EOG	O9	TP7
27	O2	O1	FCz	P6	P2
28	P6	CP3	AFz	O1	O2
29	EOG	FCz	Cz	P5	FT8
30	P5	P5	FT10	EOG	FT10
31	FT10	TP7	F5	Cz	F5
32	Fp2	O9	TP8	CPz	EOG
33	FT8	CPz	P2	F6	P5
34	O1	O2	O9	O2	CP3
35	TP8	F5	O2	TP7	FC2
36	O9	FT7	O10	CP3	FT9
37	O10	F1	F6	CP4	Fp2
38	F6	FT8	FT8	FT7	AFz
39	TP7	C5	FT9	TP8	F6

The ranking of the 39 EEG channels was calculated by the RFE method. The 17 channels over or close to motor areas of the cortex are marked with grey background for all five subjects. Underlined rank positions mark the estimated minimum of the RFE error curve for every subject from which on the error rate increases prominently (see Fig. 4 for the individual error curves)

RFE and 10-Opt proof to be capable of selecting relevant channels, whereas the FC fails for some subjects. Especially for small numbers of channels RFE is slightly superior over the FC and 10-Opt. For larger numbers of channels the performance of l0-Opt is comparable to RFE.

As can be seen in Fig. 4 it is possible to reduce the number of EEG channels significantly using the RFE method—for the

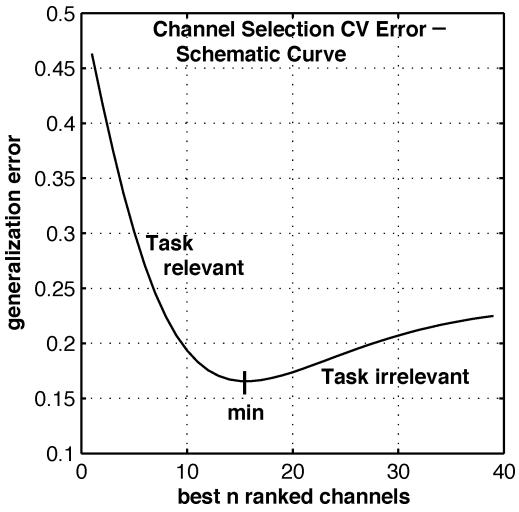


Fig. 5. Idealized generalization error curve using a channel selection method in the presence of irrelevant channels. When removing channels iteratively the classification error decreases slightly until all irrelevant channels are removed. Removing more channels results in an increase of error.

investigated experimental paradigm this can be done without a loss of classification accuracy. For example, using 8 channels for subject D yields the same error as the error obtained using all channels. On the data set of subject B the CV error of 24.5% can be reduced to 20.75% using 12 channels only.

It is not tractable to test all $(\approx 10^{11})$ possible combinations of channels to find the best combination. In this light, the 17 channels located over or close to the motor cortex can be considered a very good solution that is close to the optimal one. For rating the overall accuracy of the RFE method we, thus, trained a classifier using these 17 channels. The result averaged over the five subjects is plotted as a baseline in the last figure. The average error rate (taken over all subjects) of 24% using 12 channels is very close to the error of the baseline which is 23%.

Table I contains channel rankings, which are obtained by applying Recursive Channel Elimination to the data set of each subject.[8] As the RFE method has outperformed CF and l0-Opt, the rankings in Table I were exclusively calculated by RFE.

To interpret the table it is useful to have a closer look at Fig. 5. It shows an idealized curve for an estimate of the generalization error when using a channel or feature selection method. As we have also seen in the experiments it is possible to reduce the number of channels without a loss of accuracy. For each subject we can obtain a heuristic estimate on the number of irrelevant channels from the generalization error curves in Fig. 4. We underlined one entry in each column of Table I. The row number of that entry is an estimate for the rank position that divides task relevant channels from task irrelevant ones. For example, for subject D Fig. 4 shows a local minimum of the RFE generalization error curve at 10 channels. Thus, the best 10 selected channels can be used without increasing the error estimate.

The positions of the 17 channels over or close to the motor cortex were marked with a grey background. Except for very few of them, these channels have a high rank. For four of the subjects only few other (nonmotor) channels were ranked above the marked minimum-error positions (see underlined ranks). For

[8]Please note that in this step CV was not applied.

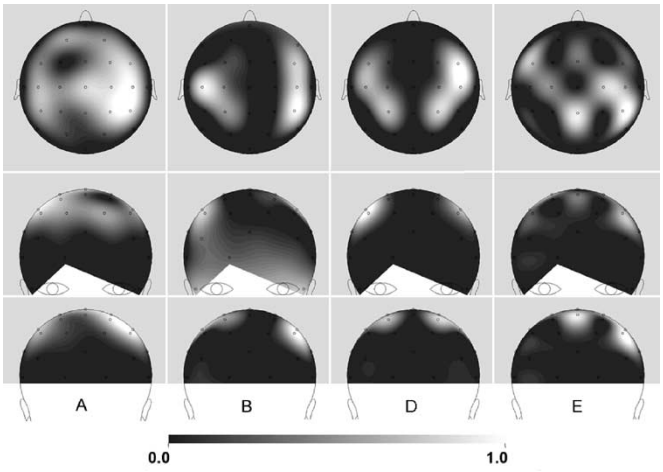


Fig. 6. Visualization of task relevant regions for subjects A, B, D and E (one subject per column) during imagined hand movements. The score for each channel was obtained by using the RFE method and is based on the full duration of 5 s. The top row depicts the view from above, the second and third row show the frontal view and view from the back. Please see also the left column of Fig. 7 for the corresponding mapping of subject C.

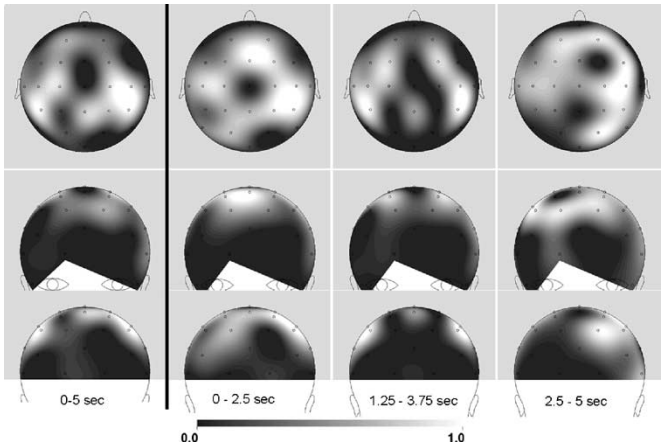


Fig. 7. Visualization of task relevant regions for subject C (top, front and back view). The leftmost column shows the scores obtained by RFE based on the complete duration of 5 s. The remaining three columns show the development of the scores over time. The rankings were obtained by applying the RFE method separately on the three shorter, overlapping time windows.

subject B channels FT9, FT10, and FP2 are relevant according to the ranking. To verify this observation we

- estimated the classification error using the seventeen motor channels and compared it to the error using the motor channels plus FT9, FT10, FP2, and EOG. Indeed by adding artefact channels the error could be reduced from 24% to 21%.
- trained an SVM based on these artefact channels only. The performance was poor: only 0.55% accuracy could be reached in a tenfold CV SVM training.[9]

That means that although feedback was not provided this subject showed task relevant muscle activity. However his performance was only supported by this muscle activity. The other four subjects did not accompany the left/right tasks with corresponding muscle movements.[10]

We conclude that the RFE method was capable of estimating physiologically meaningful EEG channels for the imagined left/right hand paradigm.

B. Visualization

The visualization of channel scores can support the analysis of BCI experiments, reveal activation patterns or channels carrying misleading artifacts and ease the choice of channel subgroups.

For visualization purposes we assigned a score calculated by RFE to each channel. The channels below the underlined entries of Table I receive a score of 0. The ones above the underlined entries are mapped to the grey value scale according to their rank. Figs. 6 and 7 show the task relevant channels for the five subjects. Black regions in both plots mark channels irrelevant for the classification task whereas white regions mark relevant ones.

[9]The ridge was explicitly optimized for this test.

[10]This observation was supported by visual inspection and frequency analysis of the raw EEG signal—only very little muscle activity or other forms of artifacts could be detected.

For all subjects the informative regions are located close to the motor cortex. Subject D shows a clear and symmetrical concentration of important channels. The second column of Fig. 6 also shows, that subject B has additional important channels outside the motor area probably resulting from muscle activity (as discussed above).

As the generalization error was minimal for the data of subject C we performed a closer examination of this data. Columns 2 to 4 of Fig. 7 visualize the spatial distribution of task specific information *over time*. We split the training data into three overlapping windows each of 2.5-s length. For every time window, we applied channel selection via RFE separately. It can be observed that the three resulting score patterns vary from window to window. This could be due to an instable channel selection. Another reason might be that the task related activation pattern changes over time. Both issues will be addressed in future experiments.

V. Conclusion

We adapted two state of the art feature selection algorithms RFE and l0-Opt as well as the FC for the special case of EEG channel selection for BCI applications.

The methods were applied to the paradigm of motor imagery. We showed that both RFE and l0-Opt are capable of significantly reducing the number of channels needed for a robust classification without an increase of error. In our experiments, the FC failed to discover satisfying channel rankings.

The reason for the decrease in performance of the l0-Opt compared to the RFE for smaller numbers of channels might be that on average the recursive l0-Opt algorithm could not decrease the number of chosen channels to less than 25 before the recursion converged. This means that all the remaining channels were ranked according to the solution of only one SVM. To overcome this shortcoming of l0-Opt we suggest the following procedure: channels are reduced with l0-Opt until the minimum l_0-norm for w is obtained. In a next step, the remaining channels are ranked using an iterative method like RFE instead of relying on a single SVM solution. This combination method was not investigated in this paper but will be subject to future research.

Although we did not incorporate explicit prior knowledge of the mental task or its underlying neural substrates, channels that are well known to be important (from a physiological point of view) were consistently selected by RFE whereas task irrelevant channels were disregarded. Furthermore the method revealed the use of muscular activity for one subject.

We introduced a method to visualize the channel rankings. This method can also be used to visualize the spatial change of task relevant information over time.

The results suggest that the RFE method can be used for new experimental paradigms in future BCI research—especially if no *a priori* knowledge about the location of important channels is available.

ACKNOWLEDGMENT

The authors would like to thank R. Rörig for her restless data processing as well as B. Battes and Prof. Dr. K. Kirschfeld for their help with the EEG recordings. They extend special thanks to G. Bakir for fruitful discussion.

REFERENCES

[1] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.

[2] B. Blankertz, G. Curio, and K. Müller, "Classifying single trial EEG: Toward brain computer interfacing," in *Advances in Neural Information Processing Systems*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2001, vol. 14.

[3] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Machine Learning Res. (Special Issue on Variable and Feature Selection)*, vol. 3, pp. 1157–1182, 2003.

[4] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *J. Machine Learning Res.*, vol. 3, pp. 1439–1461, March 2003.

[5] S. Haykin, *Adaptive Filter Theory*. Upper Saddle River, NJ: Prentice-Hall, 1996.

[6] G. Pfurtscheller, C. Neuper, A. Schlögl, and K. Lugger, "Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters," *IEEE Trans. Rehab. Eng.*, vol. 6, pp. 316–325, Mar. 1998.

[7] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: Basic principles," *Clin. Neurophysiol.*, vol. 110, no. 11, pp. 1842–1857, Nov. 1999.

[8] P. Pudil, F. Ferri, J. Novovicova, and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions," in *Proc. IEEE Int. Conf. Pattern Recognition*, 1994, pp. 279–283.

[9] B. Schölkopf and A. Smola, *Learning With Kernels*. Cambridge, MA: MIT Press, 2002.

[10] M. Schröder, M. Bogdan, W. Rosenstiel, T. Hinterberger, and N. Birbaumer, "Automated EEG feature selection for brain computer interfaces," in *Proc. 1st Int. IEEE EMBS Conf. Neural Engineering*, Mar. 2003, pp. 626–629.

[11] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[12] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, pp. 70–73, June 1967.

[13] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero-norm with linear models and kernel methods," *J. Machine Learning Res.*, vol. 3, pp. 1439–1461, Mar. 2003.

[14] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *Advances in Neural Information Processing Systems*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds. Cambridge, MA: MIT Press, 2000, vol. 12, pp. 526–532.

Thomas Navin Lal (S'96) received the diploma degree in mathematics in 2001 and spent one year with the machine learning group of Prof. Dr. T. Hofmann at Brown University, Providence, RI. He is currently a Ph.D. degree student of Prof. Dr. B. Schölkopf at the Max-Planck-Institut for Biological Cybernetics, Tübingen, Germany.

Michael Schröder received the diploma degree in computer science in 2000. He is currently a Ph.D. degree student with the Department of Computer Engineering (Prof. Rosenstiel) at the Eberhard-Karls-Universität, Tübingen, Germany.

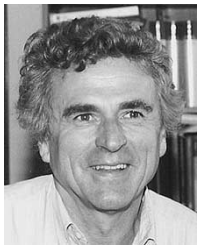
Thilo Hinterberger received the Diploma in physics from the University of Ulm, Ulm, Germany, and the Ph.D. degree in physics from the University of Tübingen, Tübingen, Germany, in 1999 on the development of a Brain-Computer-Interface, called "Thought Translation Device."

He is currently a Research Associate with the Institute of Medical Psychology and Behavioral Neurobiology at the University of Tübingen. His primary research interests focus on the further development of brain-computer interfaces and their applications but also on the development of EEG-classification methods and the investigation of neurophysiological mechanisms during the operation of a BCI using functional MRI.

Dr. Hinterberger is a member of the Society of Psychophysiological Research and the Deutsche Physikalische Gesellschaft (DPG).

Jason Weston, photograph and biography not available at the time of publication.

Martin Bogdan received the engineer diploma in signal engineering from the Fachhochschule Offenburg, Offenburg, Germany, in 1993 and the engineer diploma in industrial informatics and instrumentation from the Université Joseph Fourier Grenoble, Grenoble, France, in 1993l. In 1998, he received the Ph.D. degree in computer science (computer engineering) from the University of Tübingen, Tübingen, Germany. In 1994, he joined the Department of Computer Engineering at the University of Tübingen, where, since 2000 he heads the research group NeuroTeam. This research group deals which mainly with signal processing based on artificial neural nets and machine learning focussed on but not limited to bio-medical applications.



Niels Birbaumer was born 1945. He received the Ph.D. degrees in biological psychology, art history, and statistics from the University of Vienna, Vienna, Austria, in 1969.

In 1975-1993, he was Full Professor of Clinical and Physiological Psychology, University of Tübingen, Tübingen, Germany. In 1986-1988, he was Full Professor of Psychology, Pennsylvania State University, University Park. Since 1993, he is Professor of Medical Psychology and Behavioral Neurobiology with the Faculty of Medicine of the University of Tübingen and Professor of Clinical Psychophysiology, University of Padova, Padua, Italy. Since 2002, he is Director of the Center of Cognitive Neuroscience, University of Trento, Trento, Italy. His research topics include neuronal basis of learning and plasticity; neurophysiology and psychophysiology of pain; and neuroprosthetics and neurorehabilitation. He ha authored more than 450 publications in peer-reviewed journals and 12 books.

Among his many awards are the Leibniz-Award of the German Research Society (DFG), the Award for Research in Neuromuscular Diseases, Wilhelm-Wundt-Medal of the German Society of Psychology, Albert Einstein World Award of Science. He is President of the European Association of Behavior Therapy, a Fellow of the American Psychological Association, a Fellow of the Society of Behavioral Medicine and the American Association of Applied Psychophysiology, and a Member of the German Academy of Science and Literature.

Bernhard Schölkopf received degrees in mathematics from the University of London, London, U.K. in 1992 and in physics from Eberhard-Karls-Universität, Tübingen, Germany, in 1994, and a doctorate in computer science from the Technical University Berlin, Berlin, Germany, in 1997.

He has researched at AT&T Bell Labs, at GMD FIRST, Berlin, at the Australian National University, Canberra, and at Microsoft Research Cambridge (U.K.). He has taught at the Humboldt University and the Technical University Berlin. In July 2001, he was elected scientific member of the Max Planck Society and director at the MPI for Biological Cybernetics; in October 2002, he was appointed Honorary Professor for Machine Learning at the Technical University Berlin.

Dr. Schölkopf won the Lionel Cooper Memorial Prize of the University of London, the annual dissertation prize of the German Association for Computer Science (GI), and the prize for the best scientific project at the German National Research Center for Computer Science (GMD).