Machine Learning and Applications for Brain-Computer Interfacing

K.-R. Müller[1,2], M. Krauledat[1,2], G. Dornhege[2], G. Curio[3], and B. Blankertz[2,⋆]

[1] Technical University Berlin, Str. d. 17. Juni 135, 10 623 Berlin, Germany
[2] Fraunhofer FIRST.IDA, Kekuléstr. 7, 12 489 Berlin, Germany
[3] Dept. of Neurology, Campus Benjamin Franklin, Charité University Medicine Berlin, Hindenburgdamm 30, 12 203 Berlin, Germany
klaus@first.fraunhofer.de

Abstract. This paper discusses machine learning methods and their application to Brain-Computer Interfacing. A particular focus is placed on linear classification methods which can be applied in the BCI context. Finally, we provide an overview on the Berlin-Brain Computer Interface (BBCI).

1 Introduction

Brain-Computer Interfacing is an interesting, active and highly interdisciplinary research topic ([2,3,4,5]) at the interface between medicine, psychology, neurology, rehabilitation engineering, man-machine interaction, machine learning and signal processing. A BCI could, e.g., allow a paralyzed patient to convey her/his intentions to a computer application. From the perspective of man-machine interaction research, the communication channel from a healthy human's brain to a computer has not yet been subject to intensive exploration, however it has potential, e.g., to speed up reaction times, cf. [6] or to supply a better understanding of a human operator's mental states.

Classical BCI technology has been mainly relying on the adaptability of the human brain to biofeedback, i.e., a subject learns the mental states required to be understood by the machines, an endeavour that can take months until it reliably works [7,8].

The Berlin Brain-Computer Interface (BBCI) pursues another objective in this respect, i.e., to impose the main load of the learning task on the 'learning machine', which also holds the potential of adapting to specific tasks and changing environments given that suitable machine learning (e.g. [9]) and adaptive signal processing (e.g. [10]) algorithms are used. Short training times, however, imply the challenge that only few data samples are available for learning to characterize the individual brain states to be distinguished. In particular when dealing with few samples of data (trials of the training session) in a high-dimensional feature space (multi-channel EEG, typically several features per channel), overfitting needs to be avoided. It is in this high dimensional – small sample statistics scenario where modern machine learning can prove its strength.

⋆ The studies were partly supported by the *Bundesministerium für Bildung und Forschung* (BMBF), FKZ 01 IBE 01A/B, and by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This paper is based on excerpts of [1].

M.J. Smith, G. Salvendy (Eds.): Human Interface, Part I, HCII 2007, LNCS 4557, pp. 705–714, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The present paper introduces basic concepts of linear classification (for a discussion of nonlinear methods in the context of BCI, see [11,9,12,13,14]). Finally, we briefly describe our BBCI activities where some of the discussed machine learning ideas come to an application and conclude.

2 Linear Methods for Classification

In BCI research it is very common to use linear classifiers, but although linear classification already uses a very simple model, things *can* still go terribly wrong if the underlying assumptions do not hold, e.g. in the presence of outliers or strong noise which are situations very typically encountered in BCI data analysis. We will discuss these pitfalls and point out ways around them.

Let us first fix the notation and introduce the linear hyperplane classification model upon which we will rely mostly in the following (cf. Fig. 1, see e.g. [15]). In a BCI set-up we measure $k = 1 \ldots K$ samples $\mathbf{x}_k$, where $\mathbf{x}$ are some appropriate feature vectors in n dimensional space. In the training data we have a class label, e.g. $y_k \in \{-1, +1\}$ for each sample point $\mathbf{x}_k$. To obtain a linear hyperplane classifier

$$\mathbf{y} = \text{sign}\left(\mathbf{w}^\top \mathbf{x} + b\right) \tag{1}$$

we need to estimate the normal vector of the hyperplane $\mathbf{w}$ and a threshold b from the training data by some optimization technique [15]. On unseen data $\mathbf{x}$, i.e. in a BCI feedback session we compute the projection of the new data sample onto the direction of the normal $\mathbf{w}$ via Eq.(1), thus determining what class label $\mathbf{y}$ should be given to $\mathbf{x}$ according to our linear model.

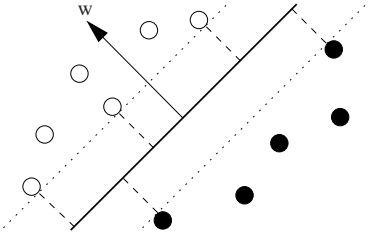

Fig. 1. Linear classifier and margins: A linear classifier is defined by a hyperplane's normal vector $\mathbf{w}$ and an offset b, i.e. the decision boundary is $\{\mathbf{x} \mid \mathbf{w}^\top \mathbf{x} + b = 0\}$ (thick line). Each of the two halfspaces defined by this hyperplane corresponds to one class, i.e. $f(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b)$. The *margin* of a linear classifier is the minimal distance of any training point to the hyperplane. In this case it is the distance between the dotted lines and the thick line. From [9].

2.1 Optimal Linear Classification: Large Margins Versus Fisher's Discriminant

Linear methods assume a linear separability of the data. We will see in the following that the optimal separating hyperplane from last section maximizes the *minimal* margin

(minmax). In contrast, Fisher's discriminant maximizes the *average* margin, i.e., the margin between the class means.

Large Margin Classification. For linearly separable data there is a vast number of possibilities to determine $(\mathbf{w}, b)$, that all classify correctly on the training set, however that vary in quality on the unseen data (test set). An advantage of the simple hyperplane classifier (in canonical form cf. [16]) is that literature (see e.g. [15,16]) tells us how to select the *optimal* classifier $\mathbf{w}$ for unseen data: it is the classifier with the largest margin $\rho = 1/\|\mathbf{w}\|_2^2$, i.e. of minimal (euclidean) norm $\|\mathbf{w}\|_2$ [16] (see also Fig. 1). Linear Support Vector Machines (SVMs) realize the large margin by determining the normal vector $\mathbf{w}$ according to

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & 1/2\,\|\mathbf{w}\|_2^2 + C/K\,\|\xi\|_1 \quad \text{subject to} \\ & y_k(\mathbf{w}^\top \mathbf{x}_k + b) \geqslant 1 - \xi_k \quad \text{and} \\ & \xi_k \geqslant 0 \qquad \text{for } k = 1,\ldots,K, \end{aligned} \tag{2}$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm: $\|\xi\|_1 = \sum |\xi_k|$. Here the elements of vector ξ are slack variables and parameter C controls the size of the margin vs. the complexity of the separation. While the user has not to care about the slack variables, it is essential to select an appropppriate value for the free parameter C for each specific data set. The process of choosing C is called model selection, see e.g. [9]. One particular strength of SVMs is that they can be turned in nonlinear classifiers in an elegant and effective way (see e.g. [16,9,12]).

Fisher's Discriminant. Fisher's discriminant computes the projection $\mathbf{w}$ differently. Under the restrictive assumption that the class distributions are (identically distributed) Gaussians of equal covariance, it can be shown to be Bayes optimal. The separability of the data is measured by two quantities: How far are the projected class means apart (should be large) and how big is the variance of the data in this direction (should be small). This can be achieved by maximizing the so-called Rayleigh coefficient of between and within class variance with respect to $\mathbf{w}$ [17,18]. The slightly stronger assumptions have been fulfilled in several of our BCI experiments e.g. in [13,14]. When the optimization to obtain (regularized) Fisher's discriminant is formulated as a mathematical programm, cf. [19,9,20], it resembles the SVM:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & 1/2\,\|\mathbf{w}\|_2^2 + C/K\,\|\xi\|_2^2 \quad \text{subject to} \\ & y_k(\mathbf{w}^\top \mathbf{x}_k + b) = 1 - \xi_k \quad \text{for } k = 1,\ldots,K. \end{aligned}$$

2.2 Some Remarks About Regularization and Non-robust Classifiers

Linear classifiers are generally more robust than their nonlinear counterparts, since they have only limited flexibility (less free parameters to tune) and are thus less prone to overfitting. Note however that in the presence of strong noise and outliers *even* linear systems can fail. In the cartoon of Fig.2 one can clearly observe that one outlier or

strong noise event can change the decision surface drastically, if the influence of single data points on learning is not limited. Although this effect can yield strongly decreased classification results for linear learning machines, it can be even more devastating for nonlinear methods. A more formal way to control one's mistrust in the available training data, is to use regularization (e.g. [21,22,15]). Regularization helps to limit (a) the influence of outliers or strong noise (e.g. to avoid Fig.2 middle), (b) the complexity of the classifier (e.g. to avoid Fig.2 right) and (c) the raggedness of the decision surface (e.g. to avoid Fig.2 right). No matter whether linear or nonlinear methods are used, one should *always* regularize, – in particular for BCI data!

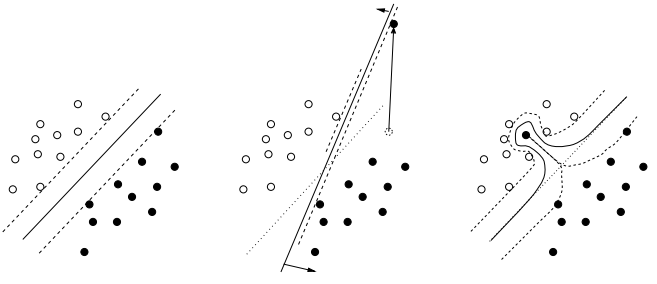

Fig. 2. The problem of finding a maximum margin "hyper-plane" on reliable data (left), data with an outlier (middle) and with a mislabeled pattern (right). The solid line shows the resulting decision line, whereas the dashed line marks the margin area. In the middle and on the right the original decision line is plotted with dots. Illustrated is the noise sensitivity: only one strong noise/outlier pattern can spoil the whole estimation of the decision line. From [23].

3 The Berlin Brain-Computer Interface

The Berlin Brain-Computer Interface is driven by the idea to shift the main burden of the learning task from the human subject to the computer under the motto 'let the machines learn'. To this end, the machine learning methods presented in the previous sections are applied to EEG data from selected BBCI paradigms: selfpaced [13,14] and imagined [24,25,26] experiments.

3.1 Bereitschaftspotential Experiments

In preparation of motor tasks, a negative readiness potential precedes the actual execution. Using multi-channel EEG recordings it has been demonstrated that several brain areas contribute to this negative shift (cf. [27,28]). In unilateral finger or hand movements the negative shift is mainly focussed on the frontal lobe in the area of the corresponding motor cortex, i.e., contralateral to the performing hand. Based on the laterality of the pre-movement potentials it is possible to discriminate multi-channel EEG recordings of upcoming left from right hand movements. Fig. 3 shows the lateralized readiness potential during a 'self-paced' experiment, as it can be revealed here by averaging over 260 trials in one subject.

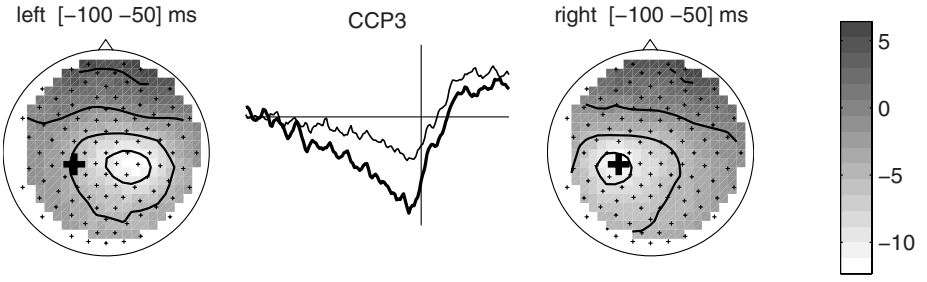

Fig. 3. The scalp plots show the topography of the electrical potentials prior to keypress with the left resp. right index finger. The plot in the middle depicts the event-related potential (ERP) for left (thin line) vs. right (thick line) index finger in the time interval -1000 to -500 ms relative to keypress at electrode position CCP3, which is marked by a bigger cross in the scalp plots. The contralateral negativation (lateralized readiness potential, LRP) is clearly observable. Approx. 260 trials per class have been averaged.

In the 'self-paced' experiments, subjects were sitting in a normal chair with fingers resting in the typing position at the computer keyboard. In a deliberate order and on their own free will (but instructed to keep a pace of approximately 2 seconds), they were pressing keys with their index and little fingers.

EEG data was recorded with 27 up to 120 electrodes, arranged in the positions of the extended 10-20 system, referenced to nasion and sampled at 1000 Hz. The data were downsampled to 100 Hz for further offline analyses. Surface EMG at both forearms was recorded to determine EMG onset. In addition, horizontal and vertical electrooculograms (EOG) were recorded to check for correlated eye movements.

In [6], it has been demonstrated that when analyzing LRP data offline with the methods detailed in the previous sections, classification accuracies of more than 90% can be reached at 110 ms before the keypress, i.e. a point in time where classification on EMG is still at chance level. These findings suggest that it is possible to use a BCI in time critical applications for an early classification and a rapid response.

Table 1 shows the classification results for one subject when comparing different machine learning methods. Clearly regularization and careful model selection are mandatory which can, e.g., be seen by comparing LDA and RLDA. Of course, regularization is of more importance the higher the dimensionality of features is. The reason of the very

Table 1. Test set error (± std) for classification at 110 ms before keystroke; ›mc‹ refers to the 56 channels over (sensori) motor cortex, ›all‹ refers to all 105 channels. The algorithms in question are Linear Discriminant Analysis (LDA), Regularized Linear Discriminant Analysis (RLDA), Linear Programming Machine (LPM), Support Vector Machine with Gaussian RBF Kernel (SVMrbf) and *k*-Nearest Neighbor (*k*-NN).

channels	LDA	RLDA	LPM	SVMrbf	*k*-NN
all	16.9±1.3	8.4±0.6	7.7±0.6	8.6±0.6	28.4±0.9
mc	9.3±0.6	6.3±0.5	7.4±0.7	6.7±0.7	22.0±0.9

bad performance of k-NN is that the underlying Euclidean metric is not appropriate for the bad signal-to-noise ratio found in EEG trials. For further details refer to [13,14]. Note that the accuracy of 90% can be maintained in recent realtime feedback experiments [29]. Here, as no trigger information is available beforehand, the classification decision is split into one classifier that decides whether a movement is being prepared and a second classifier that decides between left and right movement to come.

A similar method of classification can be applied for the discrimination of finger and arm movements, as described in [30]. This is presented in Figure 4, where a subject is sitting in front of a computer screen that shows the movement of a virtual arm which is controlled by a BCI classifier output. The classification is based entirely on pre-movement activity of the EEG, while the subject presses keys with the right index finger or lifts the arm. Scenarios like this can be used for intuitive interaction with a virtual environment.

Fig. 4. The virtual arm in the X-Rooms™virtual reality framework is an example for the applicability of slow cortical potentials in BCI scenarios. See text for details.

3.2 Motor Imagery Experiments

During imagination of a movement, a lateralized attenuation of the μ- and/or central β-rhythm can be observed localized in the corresponding motor and somatosensory cortex. Besides a usual spectral analysis, this effect can be visualized by plotting event-related desynchronization (ERD) curves [31] which show the temporal evolution of the band-power in a specified frequency band. A typical averaged ERD is shown in Fig. 5.

We performed experiments with 6 healthy subjects performing motor imagery. The subjects were sitting comfortably in a chair with their arms in a relaxed position on an arm rest. Two different sessions of data collection were provided: In both a target “L”, “R” and “F” (for left, right hand and foot movement) is presented for the duration of 3.5 seconds to the subject on a computer screen. In the first session type this is done by visualizing the letter on the middle of the screen. In the second session the left, right or lower triangle of a moving gray rhomb is colored red. For the whole length of this period, the subjects were instructed to imagine a sensorimotor sensation/movement in

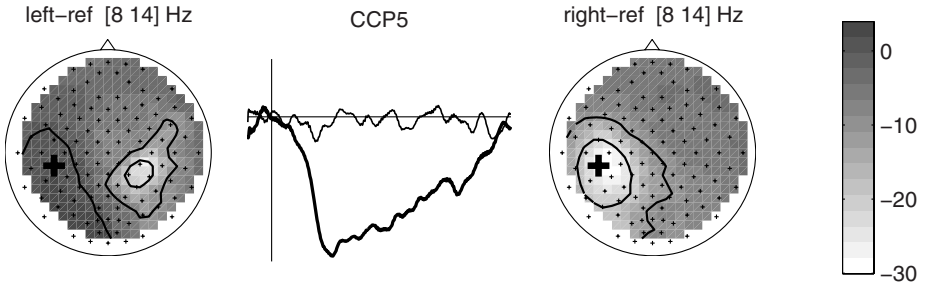

Fig. 5. This scalp plots show the topography of the band power in the frequency band 8–14 Hz relative to a reference period. The plot in the middle shows ERD curves (temporal evolution of band power) at channel CCP5 (mark by a bigger cross in the scalp plots) for left (thin line) and right (thick line) hand motor imagery. The contralateral attenuation of the μ-rhythm during motor imagery is clearly observable. For details on ERD, see [31].

Fig. 6. This screenshot presents an application of the above mentioned 1-D cursor movement. By selecting of one of the targets on the sides of the screen, it is possible to select groups of letters of the alphabet. After the first selection, the choice of letters can be further narrowed by repeated selections, until a single letter remains. This process can then be used as a text input device which is only driven by the BCI.

left hand, right hand resp. one foot. After stimulus presentation, the screen was blank for 1.5 to 2 seconds. In this manner, 35 trials per class per session were recorded. After 25 trials, there was a short break for relaxation. Four sessions (two of each training type) were performed. EEG data was recorded with 128 electrodes together with EMG from both arms and the involved foot, and EOG as described above.

An offline machine learning analysis of the "imagined"-experiments yields again high classification rates (up to 98.9% with the feature combination algorithm PROB [25,24]), which predicts the feasibility of this paradigm for online feedback situations (see also [26]). In fact, our recent online experiments have confirmed this prediction by showing high bitrates for several subjects. These subjects were untrained and had to play video games like 'brain pong', 'basket' and 'controlled 1-D cursor movement'

[32]. Depending on the 'game' scenario the best subjects could achieve information transfer rates of up to 37 Bits/min. With an accuracy of this magnitude, subjects are also able to operate text input software as presented in Figure 6. Details on this application can be found in [33].

4 Conclusion

After a brief review of general linear machine learning techniques, this paper demonstrated their application in the context of real BCI-experiments. Using these techniques, it can be seen that the paradigm shift away from subject training to individualization and adaptation ('let the machines learn') of the signal processing and classification algorithm to the specific brain 'under study' holds the key to the success of the BBCI. Being able to use (B)BCI for untrained subjects dramatically enhances and broadens the spectrum of practical applications in human-computer interfacing.

Acknowledgments. We thank our co-authors from previous publications for letting us use the figures and joint results [12,9,11,14,23].

References

1. Müller, K.-R., Krauledat, M., Dornhege, G., Jähnichen, S., Curio, G., Blankertz, B.: A note on the Berlin Brain-Computer Interface. In: Hommel, G., Huanye, S. (eds.) Human Interaction with Machines: Proceedings of the 6th International Workshop held at the Shanghai Jiao Tong University, pp. 51–60 (2006)
2. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-computer interfaces for communication and control. Clin. Neurophysiol 113, 767–791 (2002)
3. Wolpaw, J.R., Birbaumer, N., Heetderks, W.J., McFarland, D.J., Peckham, P.H., Schalk, G., Donchin, E., Quatrano, L.A., Robinson, C.J., Vaughan, T.M.: Brain-computer interface technology: A review of the first international meeting. IEEE Trans. Rehab. Eng. 8(2), 164–173 (2000)
4. Kübler, A., Kotchoubey, B., Kaiser, J., Wolpaw, J., Birbaumer, N.: Brain-computer communication: Unlocking the locked. Psychol. Bull. 127(3), 358–375 (2001)
5. Eleanor, A., Curran, E.A., Stokes, M.J.: 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, 326–336 (2003)
6. Krauledat, M., Dornhege, G., Blankertz, B., Curio, G., Mülle, K.-R.: The Berlin brain-computer interface for rapid response. Biomed. Technik 49(1), 61–62 (2004)
7. Wolpaw, J.R., McFarland, D.J., Vaughan, T.M.: Brain-computer interface research at the Wadsworth Center. IEEE Trans. Rehab. Eng. 8(2), 222–226 (2000)
8. Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kübler, A., Perelmouter, J., Taub, E., Flor, H.: A spelling device for the paralysed. Nature 398, 297–298 (1999)
9. Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B.: An introduction to kernel-based learning algorithms. IEEE Neural Networks 12(2), 181–201 (2001)
10. Haykin, S.S.: Adaptive Filter Theory. Prentice Hall, Englewood Cliffs (1995)
11. Müller, K.-R., Anderson, C.W., Birch, G.E.: Linear and non-linear methods for brain-computer interfaces. IEEE Trans. Neural Sys. Rehab. Eng. 11(2), 165–169 (2003)

12. Müller, K.-R., Krauledat, M., Dornhege, G., Curio, G., Blankertz, B.: Machine learning techniques for brain-computer interfaces. Biomed. Technik 49(1), 11–22 (2004)
13. Blankertz, B., Curio, G., Müller, K.-R.: Classifying single trial EEG: Towards brain computer interfacing. In: Diettrich, T.G., Becker, S. (eds.) Advances in Neural Inf. Proc. Systems (NIPS 01), vol. 14, pp. 157–164 (2002)
14. Blankertz, B., Dornhege, G., Schäfer, C., Krepki, R., Kohlmorgen, J., Müller, K.-R., Kunzmann, V., Losch, F., Curio, G.: Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis. IEEE Trans. Neural Sys. Rehab. Eng. 11(2), 127–131 (2003)
15. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification, 2nd edn. John Wiley & Sons, New York (2001)
16. Vapnik, V.N.: The nature of statistical learning theory. Springer, New York (1995)
17. Fisher, R.A.: he use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188 (1936)
18. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, San Diego, London (1990)
19. Mika, S., Rätsch, G., Müller, K.-R.: A mathematical programming approach to the kernel Fisher algorithm. In: Leen, T.K., Dietterich, T.G., Tresp, V. (eds.) Advances in Neural Information Processing Systems, vol. 13, pp. 591–597. MIT Press, Cambridge, MA (2001)
20. Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Smola, A., Müller, K.-R.: Constructing descriptive and discriminative non-linear features: Rayleigh coefficients in kernel feature spaces. IEEE Transaction on Pattern Analysis and Machine Intelligence 25(5), 623–628 (2003)
21. Poggio, T., Girosi, F.: Regularization algorithms for learning that are equivalent to multilayer networks. Science 247, 978–982 (1990)
22. Orr, G.B., Müller, K.-R. (eds.): Neural Networks: Tricks of the Trade. LNCS, vol. 1524. Springer, Heidelberg (1998)
23. Rätsch, G., Onoda, T., Müller, K.-R.: Soft margins for AdaBoost. Machine Learning 42(3), 287–320 (2001) also NeuroCOLT Technical Report NC-TR-1998-021
24. Dornhege, G., Blankertz, B., Curio, G., Müller, K.-R.: Increase information transfer rates in BCI by CSP extension to multi-class. In: Thrun, S., Saul, L., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems, vol. 16, pp. 733–740. MIT Press, Cambridge, MA (2004)
25. Dornhege, G., Blankertz, B., Curio, G., Müller, K.-R.: Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms. IEEE Trans. Biomed. Eng. 51(6), 993–1002 (2004)
26. Krauledat, M., Dornhege, G., Blankertz, B., Losch, F., Curio, G., Müller, K.-R.: Improving speed and accuracy of brain-computer interfaces using readiness potential features. In: Proceedings of the 26th Annual International Conference IEEE EMBS on Biomedicine, San Francisco (2004)
27. Cui, R.Q., Huter, D., Lang, W., Deecke, L.: Neuroimage of voluntary movement: topography of the Bereitschaftspotential, a 64-channel DC current source density study. Neuroimage 9(1), 124–134 (1999)
28. Lang, W., Zilch, O., Koska, C., Lindinger, G., Deecke, L.: Negative cortical DC shifts preceding and accompanying simple and complex sequential movements. Exp. Brain Res. 74(1), 99–104 (1989)
29. Krepki, R., Blankertz, B., Curio, G., Müller, K.-R.: The Berlin Brain-Computer Interface (BBCI): towards a new communication channel for online control in gaming applications (invited contribution). In: Journal of Multimedia Tools and Applications (2004)

30. Krepki, R.: Brain-Computer Interfaces: Design and Implementation of an Online BCI System of the Control in Gaming Applications and Virtual Limbs, Ph.D. thesis, Technische Universität Berlin, Fakultät IV – Elektrotechnik und Informatik (2004)
31. Pfurtscheller, G., da Silva, F.H.L.: Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin. Neurophysiol. 110(11), 1842–1857 (1999)
32. Krepki, R., Blankertz, B., Curio, G., Müller, K.-R.: The Berlin Brain-Computer Interface (BBCI): towards a new communication channel for online control of multimedia applications and computer games. In: 9th International Conference on Distributed Multimedia Systems (DMS'03), pp. 237–244 (2003)
33. Krauledat, M., Shenoy, P., Blankertz, B., Rao, R.P.N., Müller, K.-R.: Towards Brain-Computer Interfacing. In: chapter Adaptation in CSP-based BCI systems, MIT Press, Cambridge, MA (2006) (in press)