BCI Competition 2003—Data Set IIb: Support Vector Machines for the P300 Speller Paradigm

Matthias Kaper*, *Student Member, IEEE*, Peter Meinicke, Ulf Grossekathoefer, Thomas Lingner, and Helge Ritter

***Abstract*—We propose an approach to analyze data from the P300 speller paradigm using the machine-learning technique support vector machines. In a conservative classification scheme, we found the correct solution after five repetitions. While the classification within the competition is designed for offline analysis, our approach is also well-suited for a real-world online solution: It is fast, requires only 10 electrode positions and demands only a small amount of preprocessing.**

***Index Terms*—BCI competition 2003, brain-computer interface, P300 speller, SVM.**

I. Introduction

Support Vector Machines (SVM) have been successfully applied for classification and regression in various domains of pattern recognition [1]. In the present study, we utilized them for classifying electroencephalogram (EEG)-signals to detect absence or presence of the P300 component in EEG event related potentials, which is crucial for the P300 speller paradigm in Brain-Computer Interfacing [2].

The algorithm performed pretty well within the BCI-Competition 2003 [3]: It inferred the correct words and, thus, qualified as a winner. Additionally, it required the least amount of data to infer them, even in a conservative evaluation scheme.

II. Support Vector Machines

In the following, we will outline the idea of SVMs, which is discussed in more detail by, e.g., Burges [4]. An easy way to perform binary classification is to construct a hyperplane described by the weight vector $\mathbf{w}$ and the bias term b as depicted in Fig. 1. Based on a training set of l examples with the data vectors $\mathbf{x_i}$ and the corresponding class labels y_i

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l) \in R^N \times \{-1, 1\} \tag{1}$$

a machine-learning algorithm needs to find such a hyperplane according to some suitable optimality criterion. In a test phase, the class label of a new data vector $\mathbf{x}$ can be predicted by projecting $\mathbf{x}$ on the weight vector $\mathbf{w}$

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b. \tag{2}$$

Manuscript received June 30, 2003; revised December 1, 2003. This work was supported in part by the German Research Council (Deutsche Forschungsgemeinschaft) under Project SFB 360. *Asterisk indicates corresponding author.*

*M. Kaper is with the Faculty of Technology, Neuroinformatics Group, Bielefeld University, Bielefeld 33501, Germany (e-mail: mkaper@techfak.uni-bielefeld.de).

P. Meinicke is with the Bioinformatics Department at the University of Göttingen, Göttingen 37077, Germany.

U. Grossekathoefer, T. Lingner, and H. Ritter are with the Faculty of Technology, Neuroinformatics Group, Bielefeld University, Bielefeld 33501, Germany.

Digital Object Identifier 10.1109/TBME.2004.826698

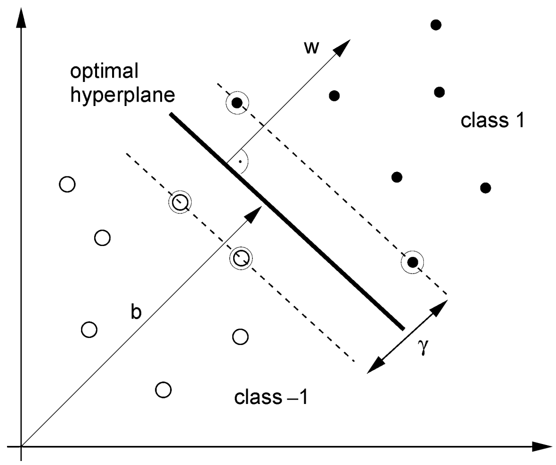

Fig. 1. SVMs find the optimal hyperplane (solid line) to separate two classes by maximizing the margin γ. It can be described by the vector $\mathbf{w}$ and the bias term b. Only support vectors (bordered circles) are necessary to calculate $\mathbf{w}$ and b.

The sign of this projection would reveal the predicted class label. While one can think of *several* possible choices of hyperplanes for dividing the data space into two subsets, as a suitable optimality criterion one may choose the *maximum margin* criterion, which favors that hyperplane with the largest separation margin γ (see Fig. 1). To describe this *optimal hyperplane*, only the vectors on the margin, the so-called *support vectors*, are necessary [5].

For a canonical representation of the hyperplane, the constraints $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$ have to be fulfilled in order to yield the margin $\gamma = 2/||\mathbf{w}||$. Maximizing the margin γ is, therefore, equivalent to minimizing $(1/2)||\mathbf{w}||^2$ subject to these constraints. Allowing violations of the constraints one may introduce the slack-variable ξ_i which gives rise to the so-called "soft-margin" SVM optimization problem

$$\begin{aligned} \min \quad & \frac{1}{2}||\mathbf{w}||^2 + C \sum_i \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i > 0 \forall i. \end{aligned} \tag{3}$$

Higher values for the regularization parameter C correspond to stronger penalties for violations. To solve this problem, it is possible to rewrite it in terms of the positive Lagrangian multipliers α_i. Then, the dual representation of (3) requires maximizing

$$L_D = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \tag{4}$$

subject to $0 \leq \alpha_i \leq \mathrm{C}$ and $\sum_i \alpha_i y_i = 0$ and yields

$$\mathbf{w} = \sum_{i}^{N_s} y_i \alpha_i \mathbf{x}_i. \tag{5}$$

N_s denotes the number of resulting support vectors. Substituting $\mathbf{w}$ in (2) with (5) results in

$$f(\mathbf{x}) = \sum_{i}^{N_s} y_i \alpha_i (\mathbf{x} \cdot \mathbf{x}_i) + b. \tag{6}$$

It can be shown [4] that the replacement of the dot product $\mathbf{x} \cdot \mathbf{x}_i$ by a positive definite symmetric Kernel function $K(\mathbf{x}, \mathbf{x}_i)$ amounts to an implicit transformation of the given data space into a (normally higher dimensional) feature space. This results in the nonlinear discriminant function

$$S(\mathbf{x}) = \sum_{i}^{N_s} y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b. \tag{7}$$

which will be used as a score in the following section. This technique leads to a more flexible decision boundary in the data space, which may increase classification accuracy. In this work, for all SVM realizations we used the Gaussian kernel

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2\sigma^2}\right). \tag{8}$$

The behavior of the SVM classifier is controlled by the regularization parameter C and the bandwidth σ of the Gaussian kernel. Both parameters need to be chosen very carefully to obtain good results.

III. Methods

Within the P300 speller paradigm, a subject is presented a 6×6 matrix, containing 36 symbols. Each row and each column is highlighted once within one trial. If the symbol, to which the subject attends, gets highlighted, a P300 component occurs in the EEG. Therefore, from 12 EEG epochs[1], a classifier should identify one row epoch and one column epoch containing a P300 to infer the attended symbol.

Consequently, from each set of six rows and six columns, we need to find one row and one column, which is most likely to be associated with a P300. For this purpose, we trained an SVM algorithm for binary classification in a training set labeled with "1" and "-1" for P300 presence/absence, and computed the value of its discriminant function (7) within a test set, with a high score indicating presence of a P300. In our case, solving the competition task consists of the following three steps.

- Data preprocessing to ease further analyses.
- SVM classifier training on the labeled training set.
- Test set classification using the trained SVM classifier to infer the words from the unlabeled test set.

A. Data Preprocessing

For each epoch, we took the amplitude values during the 600 ms after stimulus onset from the electrode positions Fz, Cz, Pz, Oz, C3, C4, P3, P4, PO7, and PO8. Afterwards, the data was bandpass filtered (0.5–30 Hz) and normalized to an interval of $[-1, 1]$.

B. SVM Classifier Training

For each trial, twelve epochs for the different rows and columns exist. Two of these epochs contain a P300, ten should not. To train the SVM, we took the two positive examples for a P300 and also two randomly chosen negative examples from each trial.

In order to obtain a good performance, the parameters C and σ need to be chosen carefully. The values of the two parameters were varied systematically and their efficiency was assessed by *cross-validation* [6] with the whole training set.

[1]An epoch corresponds to an EEG time-series accompanying a highlighting.

TABLE I
INFERRED WORDS AND ASSOCIATED ERROR RATES FROM THE EEG DATA FOR DIFFERENT REPETITION STEPS (REP)

Rep	Inferred words	Error
1	FOOD MOOT BBM PIE CAXE NCNA N5AO6 X5Z7	35.5 %
2	FOOD MOOT BBM PIE CALE TCBA Z5AOT X5Z7	29.0 %
3	FOOD MOOT HAM PIE CALE TCNA ZYAON X567	16.1 %
4	FOOD MOOT HAM PIE CALE TUNA ZYGOT 4567	3.2 %
5	FOOD MOOT HAM PIE CAKE TUNA ZYGOT 4567	0.0 %

C. Test Set Classification

To infer the symbols from the unlabeled test set, the SVM was trained on the whole training set using the optimal parameter values from fivefold cross-validation. Usually, the data is too noisy to obtain the correct symbol from only one trial. Therefore, *several* trials per symbol have to be combined. We tried to minimize this number of trials, beginning with the *first* trial for each symbol.

We regarded the value of the discriminant function (7) as a score and combined trials by adding the scores from corresponding rows/columns from different trials. This was inspired by the recently developed maximum contrast classifiers (MCC) [7], which suggest that we may interpret that score as a density difference. That row/column with the highest total score after n trials is chosen to represent the row/column with the P300. Thereby, evidence for P300 absence/presence from n trials was accumulated for each row using

$$s_i^{\text{row}} = \sum_{k=1}^{n} S\left(\mathbf{x}_{ik}^{\text{row}}\right). \tag{9}$$

Here, $S(\mathbf{x}_{ik}^{\text{row}})$ reflects the score of the epoch $\mathbf{x}_{ik}^{\text{row}}$ from the ith row of the kth trial. Afterwards, the target row r was chosen as

$$r = \arg\max_i \left(s_i^{\text{row}}\right) \tag{10}$$

with $i = 1, \ldots, 6$. This procedure was performed analogously for columns.

IV. Results

For the optimal values $C = 20.007$ and $\sigma = 27.359$, a fivefold cross-validation on the training set revealed an accuracy of 84.5% for separation of P300 from non-P300 epochs.

When analyzing the test set for the different repetitions, this resulted in the inferred symbols shown in Table I. Error rates decrease with the number of repetitions from 35.5% to 0.0%, and the correct solution was found after only five repetitions. When choosing 80% correct classification as satisfying [2], only three repetitions would be necessary.

V. Discussion

A. Performance

Using the P300 speller paradigm and the described classification method results in a very fast EEG-based BCI. In our own, more extended experiments [8], we achieved transfer rates up to 84.7 b/min with this approach.

While applying electrodes is a very time-demanding procedure, our approach has the advantage that it contents itself with data from only 10 of the 64 applied electrodes. We also tried using data from more electrode positions, but that did not improve our results. Using only a few electrodes leads to smaller data spaces, which in return results in more memory-efficient training and classification.

Another advantage of our approach is its low preprocessing requirements. No expensive mathematical operations are necessary, which makes it appropriate for an online solution. On the other hand, with an increase of computational cost, further improvements of the classification performance may be achieved. First experiments with continuous wavelets [9] have been promising.

As an important issue concerning the evaluation scheme, we like to stress that our classification results are based on the natural sequence of the experimental session, beginning with the first trial for each symbol to infer. This is in contrast to alternative schemes using arbitrary starting points, which are not very realistic in our eyes. When we are dealing with cognitive processes, it is questionable, whether the, e.g., ninth trial is comparable with the first trial. Any time-dependent processes (like changes of attention, habituation, and fatigue) are neglected when doing so. Besides, a bias of chance gets accumulated when choosing the best performing sequence from an arbitrary starting point.

B. Is the P300 Speller BCI Independent?

Further interesting findings result when investigating *what* was learned by the SVM algorithm. This can be done by analyzing the so-called *discriminative directions* from the Support Vectors of a trained SVM [10]. The amplitude values within a resulting vector reflect the importance of the specific component for classification. Recently [11], we suggested the analysis of the first principal component[2] of the discriminative directions from a Gaussian kernel SVM for further interpretation of EEG data. Applying this technique to the current data, the patterns of Fig. 2 were obtained.

Next to P300-like components, which should occur with a latency of 300 ms at parietal (and surrounding) sites, information from the parieto-occipital sites PO7 and PO8 was also utilized by the SVM for classification. This indicates that attention-dependent high-level cognitive processes may not be the sole source for the classification, but low-level sensor information from the visual cortex may contribute as well. Data from PO7 and PO8 are thereby very unlikely to reflect a P300 in this case, since P3 and P4, which are located between PO7/PO8 and Pz, do not show any evidence for such a component.

The P300 speller paradigm is in general regarded as an *independent* BCI [12], as it is said *not* to depend on activity from peripheral nerves or muscles. A deeper look at the paradigm reveals that a confoundation of eye position and attentional processes may exist, since subjects not only use *semantics* by counting the highlightings of the specific symbol, but also direct *gaze* toward that symbol. Therefore, different highlightings of focused and unfocused symbols are likely to result in different stimulations in the visual cortex, which in turn could be distinguished by a classifier.

[2]Due to its very high eigenvalue, which is also the case for the data of this paper.

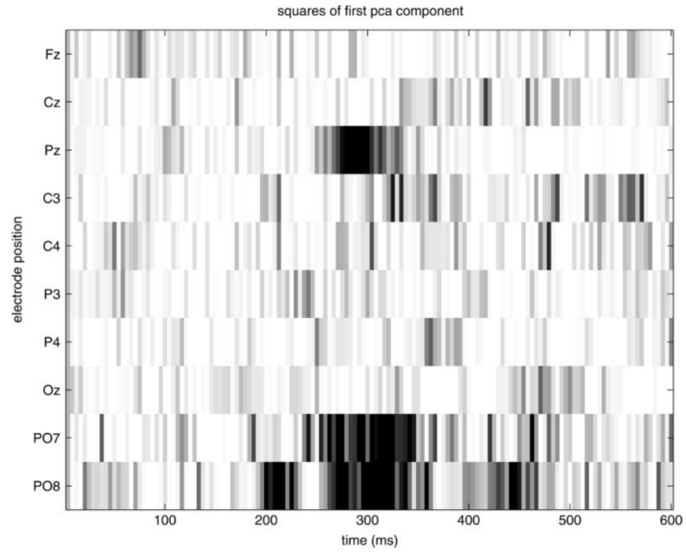

Fig. 2. Squares of the first PCA component of the SVM discriminative directions. Darker regions indicate higher amplitudes. Apparently, besides a large parietal contribution from Pz, the parieto-occipital sites PO7 and PO8 also significantly contribute to the SVM classification results.

VI. Conclusion

Applying an SVM-based method to the P300 speller paradigm resulted in perfect classification after 5 repetitions of data according to one symbol in the stimulus matrix, starting with the first trial. Applying this approach to our own experiments, we were able to achieve high transfer rates of up to 84.7 b/min [8].

Interestingly, when analyzing what the SVM algorithm learned, we found indications for significant contributions from parieto-occipital sites to the classifier output. This suggests that research community needs to investigate further the issue of full-independence in the P300 speller-paradigm.

However, within the framework set by the present contest, our algorithm inferred the correct words, it inferred them with the least amount of data, and it inferred them from the natural occurence beginning from the start. With the low amount of preprocessing and its restriction to 10 electrodes, it is furthermore very suitable for a practical realization within an online scenario.

Acknowledgment

The authors have used the SVM algorithm from the libsvm toolbox by Chang and Lin [13]. They would like to thank P. McGuire for his careful reading of the final version of the paper.

REFERENCES

[1] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Networks*, vol. 12, pp. 181–201, Feb. 2001.

[2] L. A. Farwell and E. Donchin, "Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalogr. Clin. Neurophysiol.*, vol. 70, no. S2, pp. 510–523, 1988.

[3] B. Blankertz, K.-R. Müller, G. Curio, T. M. Vaughan, G. Schalk, J. R. Wolpaw, A. Schlögl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schröder, and N. Birbaumer, "The BCI competition 2003: Progress and perspectives in detection and discrimination of EEG single trials," *IEEE Trans. Biomed. Eng.*, vol. 51, pp. 1044–1051, June 2004.

[4] C. J. C. Burges. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery* [Online], vol (2), pp. 121–167. Available: citeseer.nj.nec.com/burges98tutorial.html

[5] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

[6] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge, U.K.: Cambridge University Press, 1996.

[7] P. Meinicke, T. Twellmann, and H. Ritter, "Discriminative densities from maximum contrast estimation," in *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press, 2003.

[8] P. Meinicke, M. Kaper, F. Hoppe, M. Heumann, and H. Ritter, "Improving transfer rates in brain computer interfacing: A case study," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003.

[9] V. Bostanov, "BCI competition 2003—Data sets Ib and IIb: Feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram," *IEEE Trans. Biomed. Eng.*, vol. 51, pp. 1057–1061, June 2004.

[10] P. Golland, "Discriminative direction for kernel classifiers," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2002.

[11] M. Kaper, P. Meinicke, and H. Ritter, "Utilizing SVM's to derive psychophysiological information from a brain-computer interfacing study," in *Proc. EuroCogSci 2003*, Osnabrueck, Germany, Sept 10–13, 2003.

[12] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, pp. 767–791, 2002.

[13] C.-C. Chang and C.-J. Lin. (2001) *LIBSVM: A Library for Support Vector Machines* [Online]. Available: http://www.csie.ntu.edu.tw/\verb~cjlin/libsvm

Matthias Kaper (S'03) received the masters degrees in psychology as well as in computer science from Bielefeld University, Bielefeld, Germany, in 2001 and 2002, respectively. He is currently working towards the Ph.D. degree within the Neuroinformatics Group and the Collaborative Research Center 360 (SFB 360), in which he is working on EEG data analysis using machine learning techniques.

Peter Meinicke received the masters degree in computer science from the University of Bielefeld, Bielefeld, Germany. He received the Ph.D. degree within the graduate program "Task oriented communication."

As a Postdoc, he has been concerned with the analysis of EEG data within the Collaborative Research Center "Situated Artificial Communicators." Recently, he has joined the Bioinformatics Department at the University of Göttingen, Göttingen, Germany. His research interests are focused on the development of machine learning technology for the analysis of biological systems and signals.

Ulf Grossekathoefer is a computer science student within the Neuroinformatics Group, Bielefeld, Germany, and is concerned with EEG data analysis.

Thomas Lingner is a computer science student within the Neuroinformatics Group, Bielefeld University, Bielefeld, Germany, and is concerned with EEG data analysis.

Helge Ritter studied physics and mathematics at the Universities of Bayreuth, Heidelberg, and Munich. After receiving the Ph.D. degree in physics from the Technical University of Munich, Munich, Germany, in 1988 he visited the Laboratory of Computer Science at Helsinki University of Technology, Helsinki, Finland, and the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign.

Since 1990, he has been head of the Neuroinformatics Group at the Faculty of Technology, Bielefeld University, Bielefeld, Germany. He is primarily interested in the principles of neural computation, in particular for self-organization and learning, and their application to build intelligent systems.

In 1999, Dr. Ritter was awarded the SEL Alcatel Research Prize and in 2001 the Leibniz Prize of the German Research Foundation DFG.