A Simple Generative Model for Single-Trial EEG Classification

Jens Kohlmorgen and Benjamin Blankertz

Fraunhofer FIRST.IDA
Kekuléstr. 7, 12489 Berlin, Germany
{jek, blanker}@first.fraunhofer.de
http://ida.first.fraunhofer.de

Abstract. In this paper we present a simple and straightforward approach to the problem of single-trial classification of event-related potentials (ERP) in EEG. We exploit the well-known fact that event-related drifts in EEG potentials can well be observed if averaged over a sufficiently large number of trials. We propose to use the average signal and its variance as a generative model for each event class and use Bayes decision rule for the classification of new, unlabeled data. The method is successfully applied to a data set from the NIPS*2001 Brain-Computer Interface post-workshop competition.

1 Introduction

Automating the analysis of EEG (electro-encephalogram) is one of the most challenging problems in signal processing and machine learning research. A particularly difficult task is the analysis of event-related potentials (ERP) from individual events ('single-trial'), which recently gained increasing attention for building brain-computer interfaces. The problem is in the high inter-trial variability of the EEG signal, where the interesting quantity, e.g. a slow shift of the cortical potential, is largely hidden in the 'background' activity and only becomes evident by averaging over a large number of trials.

To approach the problem we use an EEG data set from the NIPS*2001 Brain-Computer Interface (BCI) post-workshop competition.[1] The data set consists of 516 single trials of pressing a key on a computer keyboard with fingers of either the left or right hand in a self-chosen order and timing ('self-paced key typing'). A detailed description of the experiment can be found in [1]. For each trial, the measurements from 27 Ag/AgCl electrodes are given in the interval from 1620 ms to 120 ms *before* the actual key press. The sampling rate of the chosen data set is 100 Hz, so each trial consists of a sequence of $N = 151$ data points. The task is to predict if the upcoming key press is from the left or right hand, given only the respective EEG sequence. A total of 416 trials are labeled (219 'left' events, 194 'right' events, and 3 rejected trials due to artifacts) and can be used for building a binary classifier. One hundred trials are unlabeled and make

[1] publicly available at http://newton.bme.columbia.edu/competition.htm

up the evaluation test set for the competition. It should be noted here, that we construct our classifier under the conditions of the competition, i.e. without using the test set, but since we actually have access to the true test set labels, we do not participate in the competition. In this way, however, we are able to report the test set error of our classifier in this contribution.

2 Classifier Design

As outlined in [1], the experimental set-up aims at detecting lateralized slow negative shifts of cortical potential, known as 'Bereitschaftspotential' (BP), which have been found to precede the initiation of the movement [2,3]. These shifts are typically most prominent at the lateral scalp positions C3 and C4 of the international 10-20 system, which are located over the left and right hemispherical primary motor cortex.

Fig. 1 illustrates this for the given training data set. The left panel in Fig. 1 shows the measurements from each of the two channels, C3 and C4, *averaged* over all trials for *left* finger movements, and the right panel depicts the respective averages for *right* finger movements. Respective plots are also shown for channel C2, which is located next to C4. It can be seen that, on the average, a right finger movement clearly corresponds to a preceding negative shift of the potential over the left motor cortex (C3), and a left finger movement corresponds to a negative shift of the potential over the right motor cortex (C2, C4), which in this case is even more prominent in C2 than in C4 (left panel). The crux is that this effect is largely obscured in the individual trials due to the high variance of the signal, which makes the classification of individual trials so difficult. Therefore, instead of training a classifier on the individual trials [1], we here propose to exploit the above (prior) knowledge straight away and use the averages directly as the underlying model for left and right movements.

It can be seen from Fig. 1 that the difference between the average signals C4 and C3, and likewise between C2 and C3, is decreasing for left events, but is increasing for right events. We can therefore merge the relevant information from both hemispheres into only one scalar signal by using the difference of either C4 and C3 or C2 and C3. In fact, it turned out that the best (leave-one-out) performance can be achieved when subtracting C3 from the mean of C4 and C2. That is, as a first step of pre-processing/variable selection, we just use the scalar EEG signal, $y = (C2 + C4)/2 - C3$, for our further analysis.

The respective averages, $y_L(t)$ and $y_R(t)$, of the signal $y(t)$ for all left and right events in the training set, together with the standard deviations at each time step, $\sigma_L(t)$ and $\sigma_R(t)$, are shown in Fig. 2. A scatter plot of all the training data points underlies the graphs to illustrate the high variance of the data in comparison to the feature of interest: the drift of the mean.

We now use the left and right averages and the corresponding standard deviations directly as generative models for the left or right trials. Under a Gaussian assumption, the probability of observing y at time t given the left model,

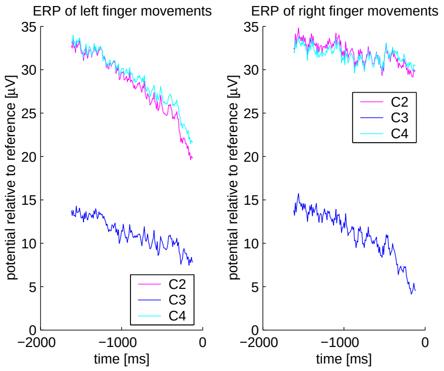

Fig. 1. Averaged EEG recordings at positions C2, C3, and C4, separately for left and right finger movements. The averaging was done over all training set trials of the BCI competition data set.

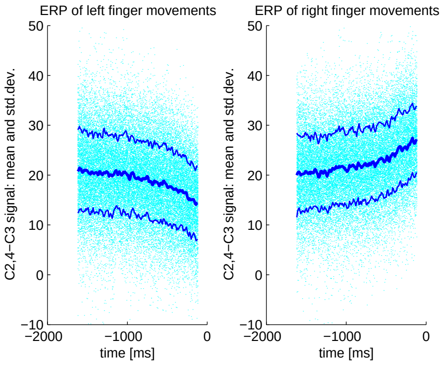

Fig. 2. Mean and standard deviation of the difference signal, $y = (C2 + C4)/2 - C3$, over a scatter plot of all data points. Clearly, there is a high variance in comparison to the drift of the mean.

$M_L = (y_L, \sigma_L)$, can be expressed as

$$p(y(t) \mid M_L) = \frac{1}{\sqrt{2\pi}\,\sigma_L(t)} \; \exp\left(-\frac{(y(t) - y_L(t))^2}{2\sigma_L(t)^2}\right). \tag{1}$$

The probability $p(y(t) \mid M_R)$ for the right model can be expressed accordingly. Assuming a Gaussian distribution is indeed justified for this data set: we estimated the distribution of the data at each time step with a kernel density estimator and consistently found a distribution very close to a Gaussian. To keep the approach tractable, we further assume that the observations $y(t)$ only depend on the mean and variance of the respective model at time t, but not on the other observations before or after time t.[2] Then, the probability of observing a complete data sequence, $\mathbf{y} = (y(1), \ldots, y(N))$, by one of the models, is given by

$$p(\mathbf{y}|M) = \prod_{t=1}^{N} p(y(t) \mid M). \tag{2}$$

Finally, the posterior probability of the model given a data sequence can be expressed by Bayes' rule,

$$p(M|\mathbf{y}) = \frac{p(\mathbf{y}|M)\; p(M)}{p(\mathbf{y})}. \tag{3}$$

We then use Bayes' decision rule, $p(M_L|\mathbf{y}) > p(M_R|\mathbf{y})$, to decide which model to choose. According to eq. (3), this can be written as

$$p(\mathbf{y}|M_L)\; p(M_L) \;>\; p(\mathbf{y}|M_R)\; p(M_R). \tag{4}$$

[2] In fact, almost all off-diagonal elements of the covariance matrix are close to zero, except for the direct neighbors, i.e. $y(t)$ and $y(t+1)$.

The evidence $p(\mathbf{y})$ vanishes in eq. (4) and we are left with the determination of the prior probabilities of the models, $p(M_L)$ and $p(M_R)$. Since there is no a priori preference for left or right finger movements in the key typing task, we can set $p(M_L) = p(M_R)$ and the decision rule simplifies to a comparison of the likelihoods $p(\mathbf{y}|M)$. If we furthermore perform the comparison in terms of the negative log-likelihood, $-\log(p)$, and neglect the leading normalization factor in eq. (1) – which turns out to not diminish the classification performance, also because the left and right standard deviations, $\sigma_L(t)$ and $\sigma_R(t)$, are very similar for this data set (cf. Fig. 2) – then the decision rule can be rewritten as

$$\sum_{t=1}^{N} \frac{(y(t) - y_L(t))^2}{\sigma_L(t)^2} \quad < \quad \sum_{t=1}^{N} \frac{(y(t) - y_R(t))^2}{\sigma_R(t)^2}. \tag{5}$$

The terms on both sides are now simply the squared distances of the respective left or right mean sequence to a given input sequence, normalized by the estimated variance of each component.

3 Results and Refinements

The above approach can readily be applied to our selected quantity y from the competition data set. The result without any further pre-processing of the signal is 20.10% misclassifications (errors) on the training set, 21.07% leave-one-out (LOO) cross-validation error (on the training set), and 16% error on the test set. Next, as a first step of pre-processing, we normalized the data of each trial to zero-mean, which significantly improved the results, yielding 14.04%/15.98%/7% training/LOO/test set error.[3] A further normalization of the data to unit-variance did not enhance the result (13.56%/15.98%/7% training/LOO/test set error).

The next improvement can easily be understood from Fig. 2. Clearly, the data points at the end of the sequence have more discriminatory power than the points at the beginning. Moreover, we presume that the points at the beginning mainly introduce undesirable noise into the decision rule. We therefore successively reduced the length of the sequence that enters into the decision rule via a new parameter D,

$$\sum_{t=D}^{N} \frac{(y(t) - y_L(t))^2}{\sigma_L(t)^2} \quad < \quad \sum_{t=D}^{N} \frac{(y(t) - y_R(t))^2}{\sigma_R(t)^2}. \tag{6}$$

Fig. 3 shows the classification result for $D = 1, \ldots, N$, $(N = 151)$, on the zero-mean data. Surprisingly, using only the last 11 or 12 data points of the EEG sequence yields the best LOO performance: the LOO error minimum (9.69%) is at $D = 140$ and 141, with a corresponding test set error of 6% in both cases.

[3] The unusual result that the test set error is just half as large as the training set error was consistently found throughout our experiments and is apparently due to a larger fraction of easy trials in the test set.

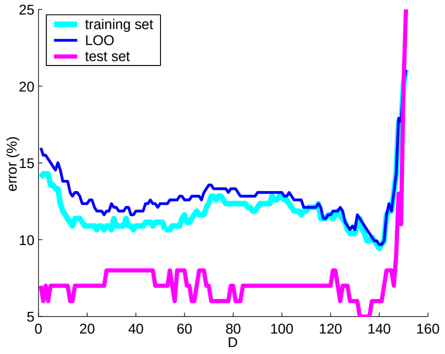

Fig. 3. Training, leave-one-out (LOO), and test set error in dependence of the starting point D of the observation window.

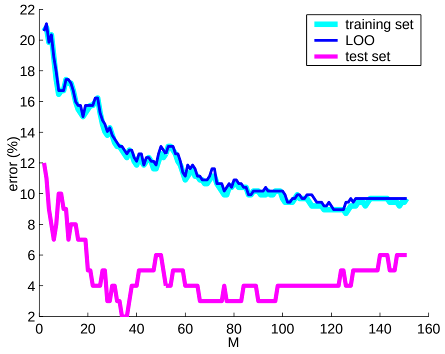

Fig. 4. Training, leave-one-out (LOO), and test set error in dependence of the size M of the zero-mean window (results for $D = 141$).

A further, somehow related improvement can be achieved by excluding a number of data points from the end of the sequence when computing the mean for the zero-mean normalization. Fig. 4 depicts the classification results for using only the first $M = 2, \ldots, N$ data points of each sequence for computing the mean for the normalization. The normalization then results in sequences that have a zero mean only for the first M data points. The LOO minimum when using $D = 141$ is at $M = 121, \ldots, 125$ (Fig. 4). The respective LOO and training set error is 8.96%. We found that this is indeed the optimal LOO error for all possible combinations of M and D. At this optimum we get a test set error of 4% for $M = 121, 122, 123$, and of 5% for $M = 124, 125$. Fig. 5 shows the respective distances (eq. (6)) of all trials to the left and right model (for $M = 121$). In Fig. 4, the test set error even reaches a minimum of 2% at $M = 34, 35, 36$, however, this solution can not be found given only the training set.

We considered other types of pre-processing or feature selection, like normalization to unit-variance with respect to a certain window, other choices of windows for zero-mean normalization, or using the bivariate C3/C4 signal instead of the difference signal. However, these variants did not result in better classification performance. Also a smoothing of the models, i.e. a smoothing of the mean and standard deviation sequences, did not yield a further improvement.

4 Summary and Discussion

We presented a simple generative model approach to the problem of single-trial classification of event-related potentials in EEG. The method requires only 2 or 3 EEG channels and the classification process is easily interpretable as a comparison with the average signal of each class. The application to a data set from the NIPS*2001 BCI competition led to further improvements of the algorithm, which finally resulted in 95–96% correct classifications on the test

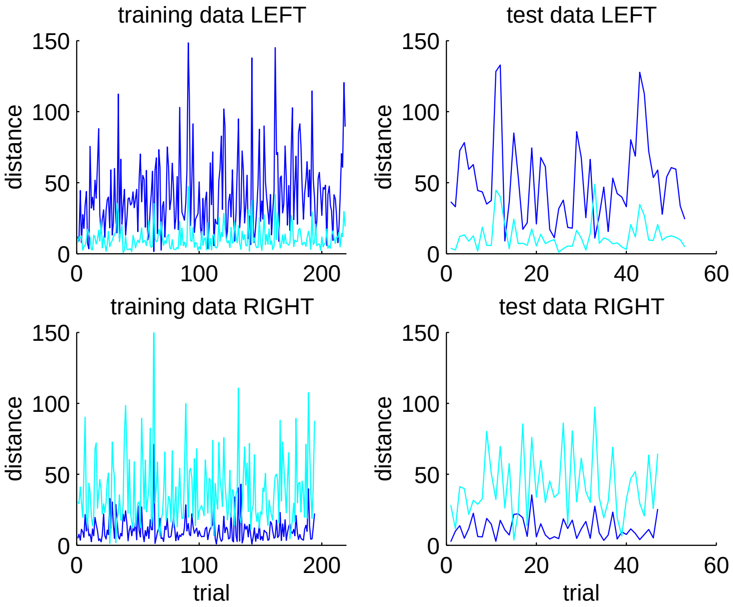

Fig. 5. Distances (cf. eq.(6)) from all trials to the finally chosen models for left and right event-related potentials ($D = 141$, $M = 121$). In almost all cases of misclassifications both models exhibit a small distance to the input. (left model: grey, right model: black)

set (without using the test set for improving the model). We demonstrated how problem-specific prior knowledge can be incorporated into the classifier design. As a result, we obtained a relatively simple classification scheme that can be used, for example, as a reference for evaluating the performance of more sophisticated, future approaches to the problem of EEG classification, in particular those from the BCI competition.

Acknowledgements: We would like to thank S. Lemm, P. Laskov, and K.-R. Müller for helpful discussions. This work was supported by grant 01IBB02A from the BMBF.

References

1. Blankertz, B., Curio, G., Müller, K.R.: Classifying single trial EEG: Towards brain computer interfacing. In Dietterich, T.G., Becker, S., Ghahramani, Z., eds.: Advances in Neural Information Processing Systems 14 (NIPS*01), Cambridge, MA, MIT Press (2002) to appear.
2. 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** (1989) 99–104
3. 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** (1999) 124–134