

Unsupervised adaptation of the LDA classifier for Brain-Computer interfaces

C. Vidaurre¹, A. Schlögl², B. Blankertz^{3,1}, M. Kawanabe¹, K.-R. Müller^{3,1}

¹Intelligent Data Analysis, Fraunhofer FIRST, Berlin, Germany

²Institute for Human Computer Interfaces, Graz University of Technology, Graz, Austria

³Department of Machine Learning, Berlin Institute of Technology, Berlin, Germany

`carmenDOTvidaurreATfirstDOTfraunhoferDOTde`

Abstract

This paper discusses simulated on-line unsupervised adaptation of the LDA classifier in order to counteract the harmful effect of non-class related non-stationarities in EEG during BCI sessions. Three types of adaptation procedures were applied to the two large BCI data sets from TU Graz and Berlin BCI project. Our results demonstrate that the unsupervised adaptive classifiers can improve performance substantially under different BCI settings. More importantly, since label information is not necessary, they are applicable to wide ranges of practical BCI tasks.

1 Introduction

A Brain Computer Interface (BCI) has to be robust against non-stationary changes [1] or adapted to these [2, 3, 4]. Some BCI users, especially at early training stages, might not generate stable EEG patterns. The system requires then supervised classifiers that can “follow” unexpected class-related changes of EEG and successfully help in the learning process [2]; however, class information is usually not available in practical BCI tasks. On the other hand, when subjects can generate stable patterns, task related EEG information might not change so drastically, but different electrode montages or task unrelated factors affect the signals. In this case, class information might not be required for the adaptation of the system. This motivated us to study whether unsupervised adaptation based on extra assumptions works in practical BCI scenarios. In BCI experiments one of the main problems from session to session or from calibration to feedback within the same session, is the bias adaptation. Typically the features move in the feature space and the classifier should be re-adjusted [2, 4, 5]. Even during a feedback session the bias must be recalculated after some time. When a subject generates almost stable patterns, one could expect the change between the vectors connecting the two class means to be small. This difference can be measured by the angle formed between the vectors connecting the mean values of each class (see Figure 1(b)).

All data processing methods used in this paper are causal and suitable for on-line and real-time realization. We describe adaptive unsupervised classifiers based on the simple and robust linear discriminant analysis (LDA). For this we exploit the fact that adapting parameters of LDA without label information is possible. The improvement in performance on two large data sets using these classifiers indicates that there exist underlying background activity that negatively affects the system performance, but that can be counteracted with the proposed methods.

2 Material and Methods

2.1 The datasets

2.1.1 Graz data

Experiments were carried out with 21 subjects without previous BCI experience. They performed motor imagery of the left and right hand to control a “basket feedback”, [6]. Each subject conducted three sessions of 9 runs and 40 trials per run (we used second and third sessions). Two bipolar channels, C3 and C4 were recorded.

2.1.2 BBCI data

We took 19 datasets recorded from 10 subjects who performed motor imagery (left-right hand, right foot) according to visual cues without feedback. The pair of tasks with best discrimination was chosen. These datasets were used because they revealed non-stationarities in a previous study [7]. Brain activity was recorded with multi-channel EEG using 55 Ag/AgCl electrodes.

2.2 Feature extraction techniques

We selected a standard choice in each laboratory: Adaptive autoregressive parameters (AAR) for the Graz data and Common spatial patterns (CSP) for the BBCI data.

2.2.1 Adaptive autoregressive parameters

To extract AAR parameters from the EEG [8], an adaptive filter based on a stable version of Recursive Least Squares (RLS) was used. AAR parameters of model order 5 were computed from two bipolar channels over C3 and C4. The logarithmic variance of the innovation process (which resulted from the adaptive filtering used) was also concatenated because it provides further information, see formula of the auto-regressive spectrum.

2.2.2 Common Spatial Patterns, CSP

CSP is a technique to analyze multichannel data based on recordings from two classes (tasks). It yields a data-driven supervised decomposition of the signal $\mathbf{x}(t)$ parameterized by a matrix \mathbf{W} that projects the signal in the original sensor space to a surrogate sensor space $\mathbf{x}_{CSP}(t)$, [9]: $\mathbf{x}_{CSP}(t) = \mathbf{x}(t) \cdot \mathbf{W}$. Each column vector of a \mathbf{W} is a spatial filter. CSP filters maximize the variance of the spatially filtered signal under one task while minimizing it for the other task. Since the variance of a band-pass filtered signal is equal to band-power, CSP analysis is applied to band-pass filtered signals to obtain an effective discrimination of mental states that are characterized by ERD/ERS effects. Detailed information about this technique can be found in [9].

2.3 Classifiers

We concentrate on a binary classification problem with linear classifiers which are specified by discriminant functions. LDA assumes the covariance matrices of both classes to be equal, Σ . We denote the means by μ_1 and μ_2 , and arbitrary feature vector by \mathbf{x} and define:

$$D(\mathbf{x}) = [b; \mathbf{w}]^T \cdot [1; \mathbf{x}] \quad (1)$$

$$\mathbf{w} = \Sigma^{-1} \cdot (\mu_2 - \mu_1) \quad (2)$$

$$\mathbf{b} = -\mathbf{w}^T \cdot \mu \quad (3)$$

$$\mu = \frac{1}{2} \cdot (\mu_1 + \mu_2) \quad (4)$$

Then $D(\mathbf{x})$ is the difference in the distance of the feature vector \mathbf{x} to the separating hyperplane described by its normal vector \mathbf{w} and the bias b . If $D(\mathbf{x})$ is greater than 0, the observation \mathbf{x} is

classified as class 2 and otherwise as class 1. Note that using a “pooled covariance matrix” instead of an averaged one does not affect the classification result. We consider five on-line adaptation schemes: two of them require label information (supervised) and the other three can update the classifier without knowing performed tasks.

2.3.1 Supervised adaptive LDA

In the supervised scenario, we can update the class means $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and the common covariance matrix $\boldsymbol{\Sigma}$ in on-line manner. LDA relies on the inverse $\boldsymbol{\Sigma}^{-1}$ of the covariance matrix $\boldsymbol{\Sigma}$ (see (2) and (3)), which can be recursively estimated applying the matrix inversion lemma, where UC is the update coefficient and $\boldsymbol{x}(t)$ is the current sample vector without the mean.

$$\boldsymbol{\Sigma}(t)^{-1} = \frac{1}{(1-UC)} \cdot \left(\boldsymbol{\Sigma}(t-1)^{-1} - \frac{1}{\frac{1-UC}{UC} + \boldsymbol{x}(t)^T \cdot \boldsymbol{v}(t)} \cdot \boldsymbol{v}(t) \cdot \boldsymbol{v}(t)^T \right) \quad (5)$$

with $\boldsymbol{v}(t) = \boldsymbol{\Sigma}(t-1)^{-1} \cdot \boldsymbol{x}(t)$. Note, the term $\boldsymbol{x}(t)^T \cdot \boldsymbol{v}(t)$ is a scalar, and no costly matrix inversion is needed. To estimate the class-specific adaptive mean $\boldsymbol{\mu}_1(t)$ and $\boldsymbol{\mu}_2(t)$ one can use:

$$\boldsymbol{\mu}_i(t) = (1-UC) \cdot \boldsymbol{\mu}_i(t-1) + UC \cdot \boldsymbol{x}(t) \quad \text{with } i := \text{class of } \boldsymbol{x}(t) \quad (6)$$

We also consider a simpler adaptive classifier (Mean classifier) which only updates the class means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ by (6), while the covariance matrix $\boldsymbol{\Sigma}$ is kept constant.

2.3.2 Unsupervised adaptive LDA I: common mean changes

As shown in [10], there are changes which affect the mean of the features. One can modify part of the bias given in (3) by adapting the common mean $\boldsymbol{\mu}(t)$ (the average of the two class means). We update the global mean $\boldsymbol{\mu}(t)$ by the same rule as (6) except that all trials from both tasks are used. This classifier is called CMean in this paper.

$$b(t) = -\boldsymbol{w}^T \cdot \boldsymbol{\mu}(t) \quad (7)$$

2.3.3 Unsupervised adaptive LDA II: common mean and covariance changes

We update the global mean and covariance matrix (CMean-CCov classifier), and keep the difference between the two class means constant. We estimate the inverse of the “pooled covariance matrix” for which no class information is needed, as in (5). We only need to subtract the common mean estimator to the current feature vector $\boldsymbol{x}(t)$. The LDA bias and weights are modified:

$$\boldsymbol{w}(t) = \boldsymbol{\Sigma}(t)^{-1} \cdot (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \quad b(t) = -\boldsymbol{w}(t)^T \cdot \boldsymbol{\mu}(t) \quad (8)$$

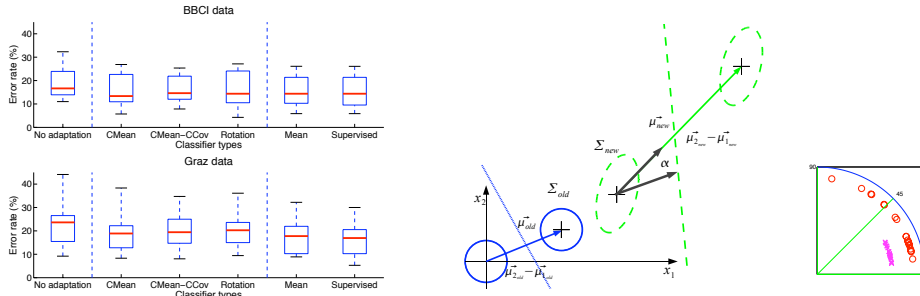
2.3.4 Unsupervised adaptive LDA III

A scaling happening in the feature space can be accounted for by using a parallel formula to that explained in [11] for the case of adaptive CSP filters:

$$\boldsymbol{w}(t) = \boldsymbol{\Sigma}(t)^{-1/2} \cdot \boldsymbol{\Sigma}^{-1/2} \cdot (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \quad b(t) = -\boldsymbol{w}^T \cdot \boldsymbol{\mu}(t) \quad (9)$$

In which one should use the “normalization assumption” of [11].

$$\boldsymbol{\Sigma}(t)^{-1/2} \cdot (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = \boldsymbol{\Sigma}^{-1/2} \cdot (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \quad (10)$$



(a) Percentiles 5, 25, 50, 75 and 95 of error rates for the two data sets. No adaptation is separated from unsupervised versions and these from the two supervised classifiers.

(b) Angle α between vectors connecting the two means, features were recorded at two different moments. (c) Angles between training and testing data.

Figure 1: Left plot: percentiles of classification error rates. Middle plot: angle formed between the vectors connecting the mean values of the two classes at two different time points. Right plot: computed angles for the two datasets.

2.3.5 Parameter selection

For each method, the update coefficient UC , the number of samples used to adapt the classifiers and the initial time for adaptation had to be selected (see [2]). All tuning parameters of "Graz data" were optimized based on the runs in session 2 for each subject and applied to session 3. The initial classifiers were calculated using the data from the previous session. For "BBCI data", the sessions were divided into two halves. The parameters were optimized in the first half and applied to the second one.

2.4 Investigating the Nonstationarities

The proposed unsupervised adaptive LDAs (Sec. 2.3.2–2.3.4) are based on the assumption that the means of the two feature distributions drift in a similar way, i.e. the difference between the two means is nearly constant. In order to quantify the validity of this assumption, the angle formed between the vectors connecting the mean values of each class of the first half and the corresponding vector of the second half is determined for each dataset.

3 Results

Figure 1(c) summarizes the results of the angles computed for every dataset. With "Graz data" the mean values were calculated using the data of session 2 for the first vector and session 3 for the second. With the "BBCI data" we used the first half against the second.

Figure 1(a) shows the percentiles 5, 25, 50, 75 and 95 of the classifiers. A significance test with a Sidak corrected p-value for multi-comparison of 1.74% revealed that no adaptation is significantly worse than all the other options. The supervised classifier was found significantly better than the rest of classifiers. Adapting the mean with and without class information did not show significant differences, although using class-labels for adapting the mean was better than CMean-CCov and Rotation classifiers. However, no differences were found between CMean and CMean-CCov, although Rotation was worse than the first one. Finally, no significant differences showed between CMean-CCov and Rotation.

Figure 2 depicts error rates of each of the classifiers versus no adaptation except in the bottom-right corner, where the mean and common mean classifiers are compared. The time in which the performance was calculated was fixed beforehand to assure causality of the results. The values below the diagonal mean that the classifier of the y-axis performs better than the one of the x-axis.

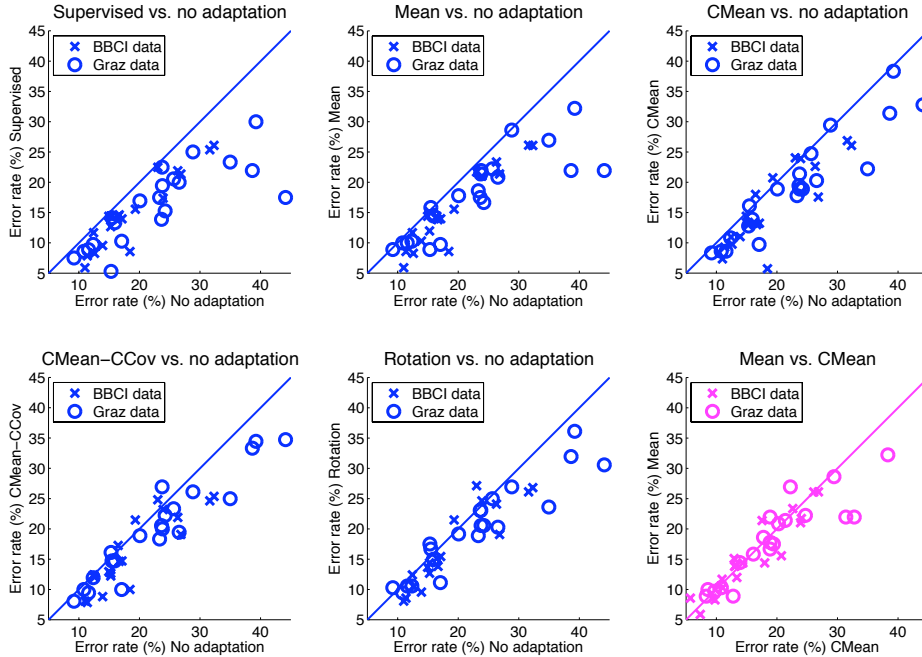


Figure 2: Comparison of classifiers based in error rates. All of them are compared to no adaptation except the bottom-right corner, in which adaptation with and without class labels are compared.

4 Discussion

Results presented in Figure 1(c) show that angles for BBCI subjects vary from 8 to 25 degrees. The features used were not adaptive and were computed with a fixed spatial filter after which band power estimates in a narrow band were calculated. Some of the subjects were naive, but did not present bigger angle-differences than experienced ones. Also, the two datasets used to estimate this difference come from the same session which would be an explanation for the small change. However, this is a realistic setting because many BCI systems record calibration and feedback runs in the same session. In contrast, all Graz subjects were inexperienced, besides the features were adaptive. These subjects show angles varying from 9 to 81 degrees.

Figures 1(a) and 2 suggest that supervised adaptation is the best option, followed by the supervised adaptation of the means. All unsupervised classifiers seem to perform very similarly, with a slight advantage of the CMean classifier, which might be due to a lower number of parameters to be adapted. It is interesting to note that adapting means with and without class-labels was not found significantly different, which is explained by the small difference between the vectors connecting the two means (small angles) found in most of the subjects. Looking at the comparison between the Mean and CMean classifiers in Figure 2 one can see that especially for 4 subjects (all of them from the Graz dataset) Mean was better than CMean. Finally, the Rotation classifier exhibits the worst average error rate of the unsupervised classifiers, and was found to be significantly worse than CMean. This might be caused because the assumptions to define problem are too strong.

5 Conclusion

When operating a BCI there is considerable fluctuation in the underlying statistics. This observation is subject and even task dependent. Compensating such non-stationary effects and investigating their underlying cause is an important mile-stone on the way to more robust BCI

systems. Our focus in this paper is to study non-task related fluctuations, for which – as we find – unsupervised data analysis methods can contribute to compensating such non-stationarity. We consider three unsupervised classifiers that are shown to successfully counteract the effect of non-class related EEG changes. These unsupervised classifiers can perform well under very different settings using CSP or AAR features for preprocessing and training and test sets from within the same session and from different ones. This is in line with the small fluctuations that can be found when analysing the change in the vectors that connect the class means at different time points. In other words, for the majority of subjects considerable signal variation is task unrelated and can thus be tackled in an unsupervised manner.

Acknowledgement

Work supported by the EU Marie Curie grant 040666 FP6, the BMBF, FKZ 01IBE01A/B, and by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, ICT-216886. This publication only reflects the authors' views.

References

- [1] B. Blankertz, M. Kawanabe, R. Tomioka, F. Hohlefeld, V. Nikulin, and K-R Müller. Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing. In *Ad. in NIPS 20*. MIT Press, Cambridge, MA, 2008. in press.
- [2] C. Vidaurre, A. Schlögl, R. Scherer, R. Cabeza, and G. Pfurtscheller. Study of on-line adaptive discriminant analysis for EEG-based brain computer interfaces. *IEEE Trans. on Biomed. Eng.*, 54:550–556, 2007.
- [3] M. Sugiyama, M. Krauledat, and K-R Müller. Covariate shift adaptation by importance weighted cross validation. *JMLR*, 8:1027–1061, 2007. accepted.
- [4] J. del R. Millán, A. Buttfield, C. Vidaurre, R. Cabeza, A. Schlögl, G Pfurtscheller, P. Shenoy, P. N. Rao, and B. Blankertz. Adaptation in brain-computer interfaces. In *Toward Brain-Computer Interfacing*, pages 303–326. MIT Press, Cambridge, MA, 2007.
- [5] J. Blumberg, J. Rickert, S. Waldert, A. Schulze-Bonhage, A. Aersten, and C. Mehring. Adaptive classification for brain computer interfaces. In *Proc 29th Ann Int Conf of the IEEE EMBS*, pages 2536–2539, 2007.
- [6] G. Krausz, R. Scherer, G. Korisek, and G. Pfurtscheller. Critical Decision-Speed and Information Transfer in the ‘Graz brain-computer-interface’. *App. Psychophysiol. and Biofeedback*, 28:233–240, 2003.
- [7] Matthias Krauledat. *Analysis of Nonstationarities in EEG signals for improving BCI performance*. PhD thesis, TU-Berlin, Fakult. IV – Elektrotech. und Inf., 2008.
- [8] Alois Schlögl. *The electroencephalogram and the adaptive autoregressive model: theory and applications*. Shaker Verlag, Aachen, Germany, 2000.
- [9] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K-R Müller. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Sign. Process. Mag.*, 25(1):41–56, January 2008.
- [10] M. Kawanabe, M. Krauledat, and B. Blankertz. A bayesian approach for adaptive BCI classification. In *Proc 3rd Int BCI Workshop 2006*, pages 54–55. Verlag TU Graz, 2006.
- [11] R. Tomioka, J. Hill, B. Blankertz, and K. Aihara. Adapting spatial filtering methods for nonstationary bcis. In *Proc. of IBIS2006*), pages 65–70, 2006.