 DOI:10.3217/978-3-85125-378-8-14

Strategies for adaptive motor imagery classification using error-related potential derived labels have unique risk profiles

Tim Zeyl[1,2] and Tom Chau[1,2]

[1] University of Toronto, Toronto, Ontario, Canada
tim.zeyl@utoronto.ca, tom.chau@utoronto.ca
[2] Bloorview Research Institute, Toronto, Ontario, Canada

Abstract

Signals measured during brain-computer interface (BCI) tasks are nonstationary, which can lead to classification errors. The error-related potential (ErrP) has been proposed for BCI error detection as well as partially-supervised classifier adaptation. We discuss how the ErrP can be incorporated into several adaptive classification methods, and the unique sensitivity that these methods have to misidentification of the ErrP. We find that the risk associated with these methods varies as a function of false positive rate for a realistic ErrP detector receiver operating characteristic and we recommend individualized biasing of the ErrP detector to account for these effects.

1 Introduction

Adaptive classification of brain-computer interfaces (BCIs) can be used to address the inherent non-stationarity of EEG data during mental tasks such as motor imagery. Class labels typically used for classifier adaptation are not available in a true BCI session, thus unsupervised adaptation has been employed as an alternative to supervised adaptation [9]. However, unsupervised methods may not be suitable when the nonstationarity affects relative class positions.

A potential compromise is to use error-related potentials (ErrPs) to generate labels for partially-supervised adaptation. Adaptation using such uncertain labels has been proposed by [5, 10], and [7] adapted an SVM classifier in the context of a code-modulated visual evoked potential speller, with benefits for participants. However, the validity of the labels depend on the accuracy of the ErrP detector, with some correct trials inevitably being interpreted as incorrect, and some incorrect trials being interpreted as correct. There is limited discussion on the risk associated with adaptation using incorrect labels and what methods are most suitable in this situation. To this end, we evaluate two adaptation methods across several ErrP detection accuracies. We assume stationary performance of the ErrP detector after the results of [2], but note that risks would increase with a nonstationary assumption.

The performance of motor imagery classifiers is dependent on choice of frequency band, and the authors of [8] showed that the most discriminative frequency changes from session to session. Therefore we build a classifier based on the filter bank common spatial pattern (FBCSP) [1] framework that uses a majority weighted vote from linear discriminant analysis (LDA)-based classifiers in each FBCSP band. This framework allows us to adapt both the ensemble weights and the base LDA classifiers separately or concurrently to either re-weight individual frequency components or change the decision boundaries in each band. In this study we evaluate the consequences of incrementally adapting these two components of the classifier at several accuracies of the ErrP detector.

 DOI:10.3217/978-3-85125-378-8-14

2 Methods

Data are taken from dataset IIb of the IVth BCI competition [4], which is comprised of 9 participants performing 5 sessions of left and right hand motor imagery, with 120-160 trials in each session. EEG is recorded with three bipolar electrodes above C3, Cz, C4 at 250 Hz with a 50 Hz notch filter. We epoch the data and apply a zero-phase filter bank of 4 Hz pass-band non-overlapping filters from $4 - 40$ Hz. In each band we extract CSP features from a 2 s window starting 1.5 s post cue. The first and last CSP features are used, which results in two features from each filter band.

We train LDA classifiers on the features from each band and combine their decisions using a weighted majority vote. Data from session 1 are used for training and the remaining sessions are used for testing. Initial weights are determined using a 10-fold cross-validation evaluating Cohen's kappa from each base classifier and normalizing the results. Then, the base LDA classifiers are retrained using all the data from session 1.

To simulate ErrPs with realistic detection accuracies, we estimated the values of the receiver operating characteristic (ROC) curves from the online ErrP detectors found in [6] and evaluated several points along the median curve, which appear in Table 1. Using these false positive rates (α) and true positive rates, we simulated ErrPs for each trial in sessions 2-4 as in [10].

We adapt both the base LDA classifiers (denoted 'BaseAdapt') and the weights of the ensemble (denoted 'Reweight') incrementally after every trial. Our estimate of the true class, $\hat{y} \in \{0, 1\}$, is derived from the classifier's output on the current trial, $\tilde{y} \in \{0, 1\}$, and our belief that an error occurred, $\tilde{E} \in \{0, 1\}$ (i.e. we detect an ErrP). Thus, $\hat{y}$ is incorrect whenever $\tilde{E}$ is incorrect. BaseAdapt updates the class means and global covariance according to the supervised LDA classifier in [9]; the 'Pmean-Gcov' unsupervised adaptive classifier from this group is included for comparison. The learning parameter, η, is set to 0.05. The Reweight strategy uses $\tilde{E}$ to implement a variant of the dynamic weighted majority of [3] that decrements the weight of incorrect experts by a factor of 0.9 and increments correct experts by a factor of 1.1. No experts are removed or created. We simulate the performance of these two adaptation strategies, as well as their combination (denoted 'Hybrid'), 50 times with independent randomly generated $\tilde{E}$ on each repetition.

False Positive Rate (α)	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
True Positive Rate	0.70	0.79	0.87	0.93	0.93	0.95	0.95	0.98

Table 1: False positive and corresponding true positive rates of the ErrP detector.

3 Results

The average classification accuracies across all 4 evaluation sessions and all participants are shown in Fig. 1a for each adaptation strategy. We see that across most values of α, on average the semi-supervised adaptation improved the accuracies over the case of no adaptation (horizontal line). Errorbar length indicates 2 standard deviations (2σ) of all 50 simulation repetitions averaged across participant and session. σ, shown as a function of α in Fig. 1b, gives a quantitative measure of the risk associated with each adaptive method. With increasing σ, there is increasing risk that the adaptation could be detrimental instead of beneficial. In general, Fig. 1 indicates that the accuracy of the BaseAdapt method decreases with increasing α, while σ increases. Even at low α, it performs no better than its unsupervised counterpart.

 DOI:10.3217/978-3-85125-378-8-14

The Reweight method has optimal accuracy and minimal σ at $\alpha \approx 0.2$. The Hybrid method obtains the best optimal average classification accuracies, but it also has the highest σ across most values of α; this is likely because it combines variability from both methods.

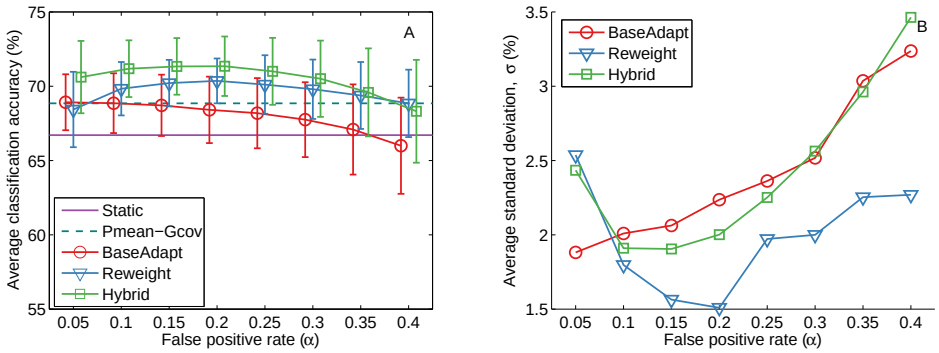

Figure 1: a. Average classification accuracies (across session and participants) as a function of false positive rate (jittered from true value for visibility). b. Standard deviation of 50 simulations as a function of false positive rate, averaged across sessions and participants.

Choosing an appropriate α for the ErrP detector involves maximizing average performance while minimizing risk, or σ equivalently. The lower quartile of the 50 simulation repetitions is a convenient measure for quantifying these two goals. The best α for each participant and each adaptation type was chosen as the one that gave the optimum lower quartile of simulation repetitions averaged across sessions. Fig. 2a compares the average classification accuracy and σ at the best α for each adaptation type and for each participant. This figure indicates highly variable performance of the adaptation strategies across participants. For some participants adaptation is not beneficial, while for others one adaptation type clearly outperforms the other. This participant dependency is also seen in the best α for each method (shown in Fig. 2b). The BaseAdapt method tends to prescribe lower α than do the Reweight and Hybrid methods where best α varies more with participant.

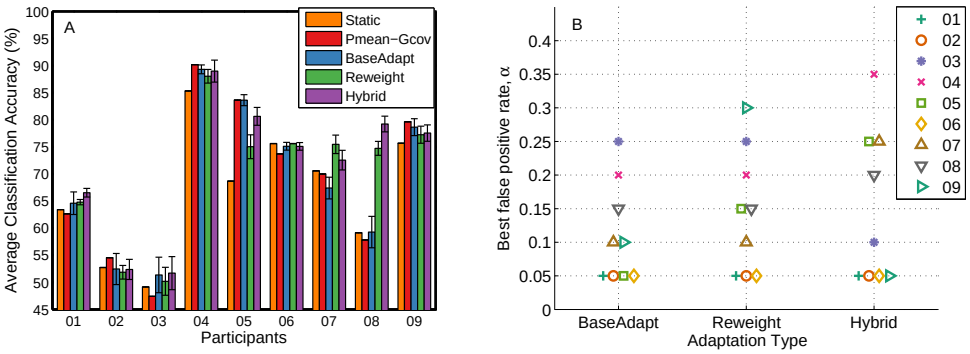

Figure 2: a. Barplot of average classification accuracies (across session) at the best false positive rate for each participant. Error bars indicate σ of 50 simulations averaged across session. b. Best false positive rate chosen for each participant (refer to legend) for each adaptation method.

4 Discussion

We find that the variance, and thus the risk associated with employing these adaptation methods changes across α in a manner unique to each method. The adaptation of the LDA base classifiers has lower risk at low α, while the re-weighting method has lower risk when α and true positive rate are balanced. The hybrid method combines the risk from both methods such that, although the average performance is the highest, it also has a high variance for most α.

LDA may be more sensitive to high α because when a false positive occurs, the error is due to a sample more likely to be further from the adapted class mean than in the case of a false negative. This drives classes closer together, which can lead to erratic movement of the class boundary as discussed in [10]. The Reweight method may not favor a single error type as such.

We found that, averaged across all participants, re-weighting the ensemble had the lowest risk. However, we find that the adaptation method with the best performance is unique to individuals, so that participant specific adaptation methods may be required. A few participants appear to benefit much more from the Reweight method compared to the BaseAdapt method. This may be due to particularly strong shifts in the most discriminative frequency for these participants. However, re-weighting may not be helpful for individuals with stationary discriminative frequency. These findings also suggest that ErrP detectors should be biased to particular α depending on the adaptation method chosen and the individual.

Future work should attempt to reduce the risk associated with adaptation using potentially uncertain labels. This may be achieved by combining the detection of an ErrP with the confidence of the BCI task classifier, or using a fuzzy classifier where weights of training samples are determined by the ErrP strength on individual trials.

References

[1] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan. Filter bank common spatial pattern (FBCSP) in brain-computer interface. In *IEEE IJCNN*, pages 2390–2397. IEEE, 2008.

[2] P. W. Ferrez and J. R. Millán. Error-related EEG potentials generated during simulated brain-computer interaction. *IEEE Transactions on Biomedical Engineering*, 55(3):923 – 929, 2008.

[3] J. Z. Kolter and M. A. Maloof. Dynamic weighted majority: An ensemble method for drifting concepts. *J Mach Learn Res*, 8(12):2755 – 2790, 2007.

[4] R. Leeb, F. Lee, C. Keinrath, R. Scherer, H. Bischof, and G. Pfurtscheller. Brain–computer communication: motivation, aim, and impact of exploring a virtual apartment. *IEEE T Neur Sys Reh*, 15(4):473–482, 2007.

[5] A. Llera, M. van Gerven, V. Gómez, O. Jensen, and H. Kappen. On the use of interaction error potentials for adaptive brain computer interfaces. *Neural Networks*, 24(10):1120–1127, 2011.

[6] N. M. Schmidt, B. Blankertz, and M. S. Treder. Online detection of error-related potentials boosts the performance of mental typewriters. *BMC Neurosci*, 13:19, Feb. 2012.

[7] M. Spüler, W. Rosenstiel, and M. Bogdan. Online adaptation of a c-VEP brain-computer interface(BCI) based on error-related potentials and unsupervised learning. *PLOS ONE*, 7(12):e51077, Dec. 2012.

[8] K. P. Thomas, C. Guan, C. T. Lau, A. P. Vinod, and K. K. Ang. Adaptive tracking of discriminative frequency components in electroencephalograms for a robust brain–computer interface. *J Neural Eng*, 8(3):036007, 2011.

[9] C. Vidaurre, M. Kawanabe, P. von Bünau, B. Blankertz, and K. R. Müller. Toward unsupervised adaptation of LDA for brain–computer interfaces. *IEEE T Biomed Eng*, 58(3):587–597, 2011.

[10] T. J. Zeyl and T. Chau. A case study of linear classifiers adapted using imperfect labels derived from human event-related potentials. *Pattern Recogn Lett*, 37(0):54–62, Feb. 2014.