Optimal Session-to-Session Transport for BCI

Steffen Schneider
Technical University of Munich, Munich, Germany
steffen.schneider@tum.de

Abstract

In this report, we present and discuss a pipeline to classifier error-related potentials (ErrPs) in electroencephalography (EEG) recordings during a cursor movement task. We compare several approaches for preprocessing, feature selection and classification. Maybe surprisingly, good performance is obtained even with simple feature selection methods apply directly to the time traces. We obtain an ACC of XX %, F1-score %, AUC % using a final ensemble of Random Forest and SVM classifiers[1].

1 Introduction

Classification of error-related potentials (ErrPs) is a common requirement in the design of brain-computer interfaces (BCIs) [8, 6, 4]. ErrPs are evoked by an error event, which can be either induced by the experiment or conducted by the user. Based on the kind of error, ErrPs exhibit characteristic properties [9], most notably a low-frequency signal decrease 200ms after the stimulus (N200), followed by a peak 300ms after the stimulus (P300). As one application, deoded ErrPs is useful for direct feedback loops between the user in the BCI, possibly for enhancing performance of the BCI itself or to learn behaviour desirable for the user [4].

In this report, we present and discuss a pipeline to classifier error-related potentials (ErrPs) in electroencephalography (EEG) recordings during a cursor movement task.

2 Methods

In this section, a brief overview of the used methods is given. It should be noted that mostly simple features were used. More sophisticated schemes such as the application of deep learning algorithms [10] are likely to fall short unless pre-trained models are available. The main focus on this section is on the domain adaptation setting present in the dataset. Domain adaption via Adaptive Subspace Feature Matching has been evaluated previously [1]. In this work, we will consider the use of optimal transport [3, 5, 2] for session-to-session transfer in BCI. As no labeled data for the target domain exists, an additional emphasis will be on the visual inspection of classiciation results. The code used to generate the results presented here is available at `github.com/stes/bci`.

2.1 Feature Selection

To determine the relevance of different time points, the t-statistic between the two classes is computed. From this, the channel and timesteps of highest significance, i.e., with lowest p-value, is selected. [2]

[1] Values will be filled in after evaluation on the non-public testset.

[2] Already at this point, we shall note that the t-test is not designed to compare p-values. This is why the p-values here are of no statistical meaning of interest and should be considered as a mean to perform feature selection.

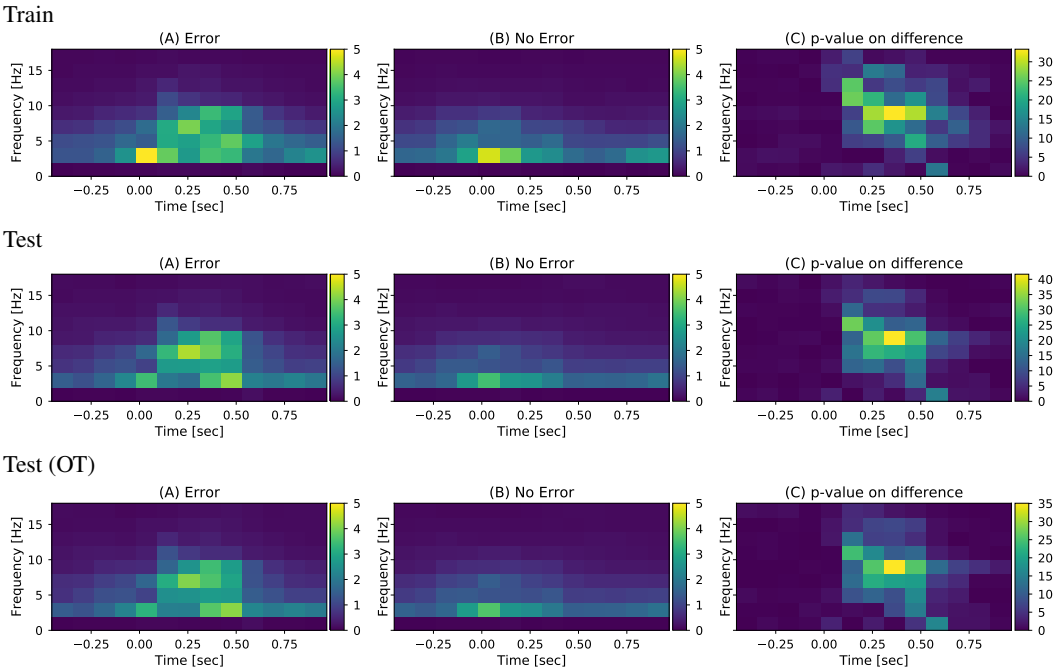

Figure 1: Spectrogram analysis of the dataset. (A) and (B) depict the mean spectograms for the Cz channel during Error and NoError events, respectively. To evaluate discriminative power, a p-test is performed between the samples, yielding the distribution depicted in (C) for the Cz channel.

Peak Picking Choose maximum peak around the identified time point of significance. An 80ms window was used. The tests were carried out in the following three predefined time segments after the stimulus

- N200: $[180\text{ms}, 210\text{ms}]$
- P300: $[250\text{ms}, 350\text{ms}]$
- delayed P300: $[350\text{ms}, 450\text{ms}]$

After obtaining the timepoint with minimal p-value in each region, the top n channels were selected for each timepoint using the same criterion. Finally, the maximum value from a predefined region around the timepoints was extracted for each channel, yielding a feature vector $z \in \mathbb{R}^{N\times 9}$ for $n = 3$.

Spectogram analysis The spectrogram was computed in the range of 0-20 Hz. In the further analysis, it is only considered as a visual sanity check as performance was inferior to the peak picking method.

Domain Adaptation by Optimal Transport In this work, we adapt the notion of domain adaptation also given in the review by [7]. A domain is denoted as a tuple $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ of a space $\mathcal{X}$ and a distribution $\mathbb{P}_{\mathcal{X}}$. In our setting, we do not only deal with a single source and a single target domain, but rather with a set of domains $\{\mathcal{X}^k\}_{k=1}^N$ which are part of a common signal space $\mathcal{U} \supset \mathcal{X}^k \; \forall k \in [N]$. We consider each domain here to be fully defined as a set of samples directly given by $\mathcal{X}^k := \{\mathbf{x}^{(j)}\}_{j=1}^{N_k}$ drawn *i.i.d.* from $\mathbb{P}_{\mathcal{X}}$.

As a last requirement, we assume that a feature space $\mathcal{V}$ and *measurement functions* $\Phi^k : \mathcal{V} \mapsto X^k$ exists such that for each $\mathbf{x} \in \mathcal{X}^k$, there exists $v \in \mathcal{V}$ with $\Phi^k(v) = \mathbf{x}^k$. This view can also be extended in a probabilistic way by adding noise to the measurement process. For a subset of domains with $k \in \mathcal{I}$, labels $\mathcal{Y}^k = \{y^{(j)}\}_{j=1}^{N_k}$ are available. Goal of the adaptation is to be able to apply an algorithm fitted to $(\mathcal{X}^{\mathcal{I}}, \mathcal{Y}^{\mathcal{I}})$ on all data domains by transforming $\mathbf{x} \in \mathcal{X}^k$ to $\mathbf{x}' \in \mathcal{X}^{\mathcal{I}}$ such that the latent representation $v \in \mathcal{V}$ (the content) of both samples is preserved.

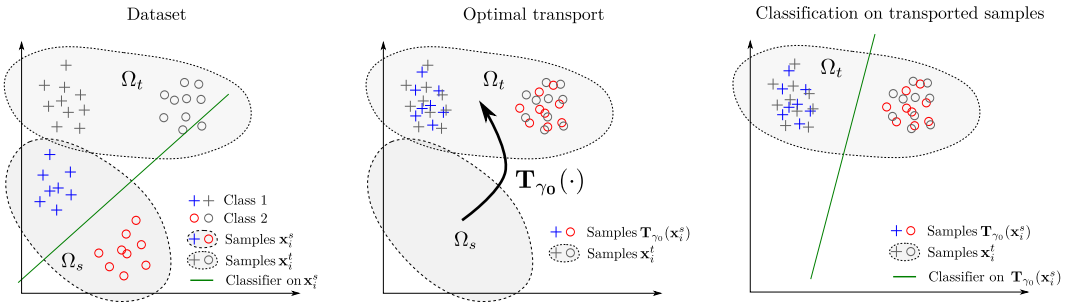

Figure 2: Schematic overview of optimal transport. Given the source and target domains Ω_s and Ω_t, optimal transport is used to compute a mapping T between source and target domain, transforming the training dataset to match the target distribution. Finally, the classifier is fitted to $T(\Omega_s)$, making it more suitable for classifying the data from Ω_t. In the context of BCI research, we propose the use of this technique between features from different sessions or subjects. Reproduced from [2].

Recently [2] presented a methods to apply mechanisms from optimal transport [3, 5] to this domain adaptation problem. We evaluate Optimal Transport for session-to-session transfer in the detection of ErrPs. The general scheme of transferring a source to the target domain using label information is depicted in figure 2.

2.2 Classification

For robust classification, we constructed an ensemble classifier to obtain a measure of uncertainty over the predictions. Candidates for the ensemble are Random Forest Classifiers (RFOs), Support Vector Machines (SVMs) with variable hyperparameters and a Linear Discriminant Analysis (LDA) without tunable hyperparameters.

For the final evaluation, a 10-fold cross validation is used and evaluated 10 times with random dataset shuffles.

Resolving Disagreement For the final application, a model ensemble trained on all training dataset splits is used to estimate uncertainty of the estimate. In this process, samples with less then a fraction of ϵ or more than $1 - \epsilon$ of all classifiers reporting the Error class are re-evaluated by a resolver taking into account more datapoints than the previous feature extraction scheme.

Using the samples already assigned to one of the classes, two templates k_0 and k_1 are computed as the average of these samples. A simliarity metric followed by softmax normalization is used to obtain a score $p \in [0, 1]$ for the data sample x, yielding

$$p = \frac{\exp(x^T k_1)}{\exp(x^T k_0)\exp(x^T k_1)}. \tag{1}$$

A threshold of $\theta = .8$ obtained by cross-validation is used to discriminate between error and non-error classes.

2.3 Visualization

As labels are not available for the validation set, we offer a qualitative evaluation of the classifier performance. In figure 1, we use the spectograms as well as significant differences between time/frequency bins as one indicator. As a second method depicted in figure 3, overview images at the extracted time points are given in electrode space as well as the according time traces for channels with highest significance.

Fold	LDA	LinSVM	RFO
0	87.2 %	85.7 %	85.7 %
1	86.8 %	86.2 %	84.8 %
2	88.3 %	88.3 %	88.0 %
3	85.7 %	85.8 %	84.7 %
4	87.5 %	87.8 %	85.7 %

	Ensemble	Resolver	OT Ensemble	OT Resolver
0	86.7 %	90.3 %	100.0 %	100.0 %
1	89.0 %	90.0 %	100.0 %	100.0 %
2	81.7 %	76.0 %	100.0 %	98.0 %
3	85.7 %	81.0 %	100.0 %	98.7 %
4	88.0 %	87.3 %	100.0 %	99.3 %
5	89.7 %	92.3 %	100.0 %	98.7 %
6	81.7 %	89.3 %	100.0 %	100.0 %
7	91.0 %	91.7 %	100.0 %	99.3 %
8	85.3 %	88.0 %	100.0 %	100.0 %
9	95.7 %	97.0 %	100.0 %	100.0 %

Table 1: *left:* Single classifier performance with best hyperparameter values. *right:* Classification accuracy for ensemble classifiers without and with Optimal Transport feature preprocessing.After the ensemble, a resolver is applied to all classes with an ensemble probability in $[\epsilon, 1-\epsilon]$.

3 Experiments

Two datasets were considered in the initial phase of this work. In addition to the original dataset, the dataset from [9] was pre-processed as well, but not included in the evaluation presented in this report. We evaluate the approach on the ICS ERP Dataset, which is comprised of 300 training and 300 validation epochs with approx. 35% of samples representing error events.

3.1 Feature Selection

Feature Selection was performed as described in the Methods section. In the overview figures 3, a qualitative comparison between the ground truth dataset split and the one obtained on the test set is given.

Especially during the P300 response, highest differences (as expected) are observed in the Cz electrode. During the N200 response and the later response denoted as P400 (which is probably a delated P300), we found the Pz electrode to offer the best discriminative power.

As one qualitative measure of test performance, we compute the same statistics for the test set and compare these timesteps. While the N200 and P400 responses are very consistent over datasets, a slight shift could be observed for the exact P300 timepoints.

3.2 Classifiers

Given the limited amount of data samples, we evaluated the use of SVMs and Random Forests as the classifiers. Evaluation using decompositions methods such as principial component analysis over a larger amount of feature revealed that this sort of feature preprocesing is not really need. Rather, the results in table 1 computed on only 9 features extracted directly from the signal traces is sufficient to obtain a reasonable performance on the training dataset.

Quite consistently, we found that linear SVMs outperform RFOs and RBF SVMS in classification performance. Below, we report the cross-validation results for the approaches:

- RFO: The number of tree ensembles was varied between 2 and 20. After cross-validation, the value was fixed to 18
- Linear SVM: The regularization parameter C was varied from 10^{-6} to 10^{2}. Best performance was obtained for $10^{-1} \ldots 1.5$.
- RBF SVM: The regulariziation parameter was varied in the same range as for the linear SVM. However, no satisfying performance was obtained for the simple feature set, which is why this method was dismissed after cross validation
- LDA offered consistent results, which is why we included it in the final ensemble

After fixing hyperparameters, a final ensemble was trained, once on the original features, once on the features adapted to test space by the optimal transport algorithm.

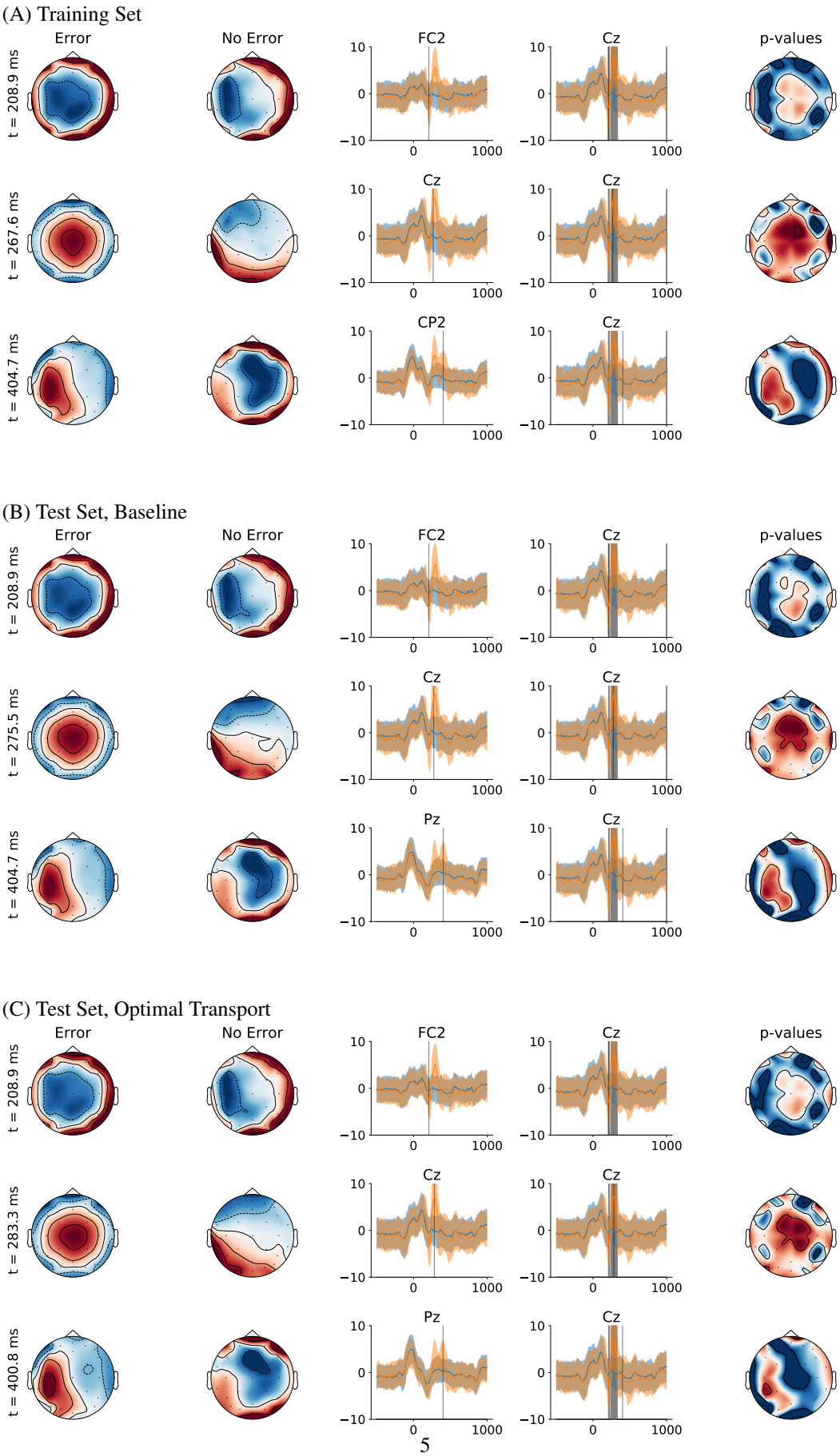

Figure 3: Qualitative comparison of ERPs obtained from splittig the test dataset with a classifier. The same statistics used for feature selection in the training set (A) are computed for the test set obtained by the baseline classifer (B) as well as the classifier trained on optimal transport samples (C).

The results on the training set can be found in table 1. While the OT ensemble performed (suspiciously?) well during the cross-validation, it will be interesting to evaluate both approaches using labels from the second session.

4 Discussion

Given the small size of the dataset, interpretation of performance and generalization of the results presented here should be done with care. In general, it was shown that with a minimal amount of features obtained by simple statistical measures, a decent classification performance on the test set is achievable. Overfitting was prevented by extensive use of model ensembles. To overcome the issue that in some cases, the selected features might not be sufficient for a decision, an additional method for resolving disagreement was presented.

Considering Session-to-Session transfer, a classification pipeline using optimal transport from the source domain (first session) to the target domain (second session) was implemented. While performance on the training set was boosted significantly, the performance on the test set will be interesting to consider.

References

[1] X. Chai, Q. Wang, Y. Zhao, Y. Li, D. Liu, X. Liu, and O. Bai. A Fast, Efficient Domain Adaptation Technique for Cross-Domain Electroencephalography(EEG)-Based Emotion Recognition. *Sensors*, 17(5):1014, may 2017.

[2] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal Transport for Domain Adaptation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, X(X):1–14, 2015.

[3] M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. 2013.

[4] S. Ehrlich and G. Cheng. A neuro-based method for detecting context-dependent erroneous robot action. *IEEE-RAS International Conference on Humanoid Robots*, pages 477–482, 2016.

[5] R. Flamary, M. Cuturi, N. Courty, and A. Rakotomamonjy. Wasserstein Discriminant Analysis. 2016.

[6] M. A. Lebedev and M. A. L. Nicolelis. Brain-machine interfaces: past, present and future. *TRENDS in Neurosciences*, 29(9):536–546, 2006.

[7] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[8] S. Silvoni, A. Ramos-Murguialday, M. Cavinato, C. Volpato, G. Cisotto, A. Turolla, F. Piccione, and N. Birbaumer. Brain-computer interface in stroke: a review of progress. *Clinical EEG and Neuroscience*, 42(4):245–252, 2011.

[9] M. Spüler and C. Niethammer. Error-related potentials during continuous feedback: using EEG to detect errors of different type and severity. *Frontiers in human neuroscience*, 9:155, 2015.

[10] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn. Deep Feature Learning for EEG Recordings. 2015.