When brain and behavior disagree: A novel ML approach for handling systematic label noise in EEG data

Nico Görnitz*,[1], Anne K. Porbadnigk*,[1], Marius Kloft[2], Alexander Binder[1], Claudia Sannelli[1], Mikio Braun[1], and Klaus-Robert Müller[1,3]

[1] Machine Learning Lab, Berlin Institute of Technology, Berlin, Germany
[2] Courant Institute of Mathematical Sciences, New York, and Memorial Sloan-Kettering Cancer Center, New York, NY, USA
[3] Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea
* Authors contributed equally

Abstract. Neuroscientific data is typically analyzed based on the behavioral response of the participant. However, the behavioral errors made may or may not be in line with the neural processing. In particular in experiments with time pressure or studies where the threshold of perception is measured, the error distribution deviates from uniformity due to the heteroscedastic nature of the underlying experimental set-up. When we base our analysis on the behavioral labels as usually done, then we ignore this problem of systematic and structured (non-uniform) label noise and are likely to arrive at wrong conclusions in our data analysis. This paper contributes a remedy to this important scenario: we present a novel approach for a) measuring label noise and b) removing structured label noise. We show its usefulness for an EEG data set recorded during a standard d2 test for visual attention.

1 Introduction

Each trial in a neuroscientific experiment is typically associated with the stimulus that was shown to the participant and his/her response to it, e.g. a button press. Typically, stimulus or behavioral response are used as labels and the neural data is analyzed by averaging trials accordingly, in categories such as 'target' vs. 'non-target' (stimulus) or 'Yes' vs. 'No' (response). However, while the conventional approach assumes brain and behavior to be in line with each other, they might disagree. For example, this can be the case for tasks with stimuli at the threshold of perception (non-conscious processing) and experiments with time pressure, resulting in responses that are unreliable or even close to random guessing. The effect may be exacerbated when participants become distracted, bored, or sleepy, resulting in a significant increase of mislabeled trials (see also [12]). We assume this label noise to be sytematic. This, in turn, challenges most of the learning algorithm employed today, which struggle not only with non-uniform label noise [1], but also with highly disbalanced classes (e.g., more false than correct responses in complex tasks), and the presence of additional brain states (e.g., 'participant tired').

As a remedy, we propose an unsupervised learning algorithm called Latent Variable Support Vector Data Description (LATENTSVDD) to tackle this challenge, focusing

on electroencephalography (EEG) data. LATENTSVDD is an extension of SVDD [13] which itself is an unsupervised anomaly detection method. The main idea is to introduce latent variables into the former which can be understood as different brain states. We show the usefulness of this new framework on EEG data from a d2 attention test. Specifically, we aim at determining whether a participant has processed a potential error on a *neural* level. Neurophysiologically, response errors are accompanied by two components in the event-related potentials (ERPs): the error negativity (N_e) and the error positivity (P_e). The N_e has been attributed to the comparison process rather than its outcome, while the P_e has been suggested to be related to error or post-error processing [5]. Therefore, we focus on the P_e in the following, which is characterized by a centro-parietal maximum 200–500ms after key stroke [9,4,6,7].

2 Learning Methodology

We consider a learning scenario where we have varying confidence in the labels. As a remedy, we propose a measure based on kernel target alignment scores (KTA) and a data-driven learning approach (LATENTSVDD) for tackling the following problems: (1) detecting anomalous trials, (2) handling systematic label noise, (3) revealing latent (brain) states, (4) verifying the results.

2.1 Kernel Target Alignment (KTA)

We are given N labels $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N \in \{+1, -1\}$ and a Gram matrix $K \in M(N \times N, \mathbb{R})$. Kernel target alignment (KTA) [3] is a method to measure the fit between the gram matrix and the label set. A high value is achieved, if data points of one class lie nearby and data points of opposite classes are far away. Mathematically, it is defined as:

$$\text{KTA}(K, \boldsymbol{y}) = \langle K, \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} \rangle_F / \sqrt{\langle K, K \rangle_F N^2}$$

Since we cannot access the underlying ground truth of an EEG experiment, KTA scores are useful as a natural indicator for the fit between labels and data before and after de-noising.

2.2 Latent Variable Support Vector Data Description (LATENTSVDD)

Our approach is based on the paradigms of support vector learning [15,10], density level set estimation, support vector data description (SVDD) [11,13] and extensions [8]. We are given N data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, where $\boldsymbol{x}_i$ which lie in some input space $\mathbb{R}^d$. The data is usually mapped from the input space into some feature space $\phi \colon \mathbb{R}^d \to \mathcal{F}_c$.

In SVDD, the goal is to find a model $f \colon \mathbb{R}^d \to \mathbb{R}$ and a density level-set $D_R = \{\boldsymbol{x} \colon f(\boldsymbol{x}) \leq R^2\}$ containing most of the normal data. In case of the SVDD, $f_{\text{SVDD}}(\boldsymbol{x}) = ||\boldsymbol{c} - \phi(\boldsymbol{x})||^2$ and parameter estimation corresponds to solving:

$$\{\boldsymbol{c}_{\text{SVDD}}, \boldsymbol{\xi}_{\text{SVDD}}, R_{\text{SVDD}}\} = \underset{\boldsymbol{c}, \boldsymbol{\xi} \geq 0, R \geq 0}{\operatorname{argmin}} \; R^2 + C \sum_{i=1}^{n} \xi_i \,, \quad \text{s.t. } ||\boldsymbol{c} - \phi(\boldsymbol{x}_i)||^2 \leq R^2 + \xi_i \quad \forall\, i$$

In this paper, we extend the classical mapping f_{SVDD} by inclusion of a latent variable $\boldsymbol{z} \in \mathcal{Z}$ in an joint feature map $\Psi : \mathbb{R}^d \times \mathcal{Z} \to \mathcal{F}$. As an consequence, the resulting model $f : \mathbb{R}^d \to \mathbb{R},\ \boldsymbol{x} \mapsto \min_{\boldsymbol{z} \in \mathcal{Z}} ||\boldsymbol{c} - \Psi(\boldsymbol{x}, \boldsymbol{z})||^2$ becomes more expressive (see also [14]). The latent state variable of a given data point $\boldsymbol{x}$ can be inferred by $g(\boldsymbol{x}) = \operatorname{argmin}_{\boldsymbol{z} \in \mathcal{Z}} ||\Psi(\boldsymbol{x}, \boldsymbol{z})||^2 - 2\langle \boldsymbol{c}, \Psi(\boldsymbol{x}, \boldsymbol{z}) \rangle$. The resulting model, we call LATENTSVDD.

We define our joint feature map as a variant of the multi-class joint feature map [14] $\Psi(\boldsymbol{x}, \boldsymbol{z}) = \phi(\boldsymbol{x}) \otimes \delta(\boldsymbol{z}_k, \boldsymbol{z})$ with $k \in \{1, \ldots, 12\}$ which is more than we expect. We train our method on all available data points, which delivers anomaly scores and latent variables for each of them. Labels are assigned depending on a majority vote for every latent class.

We designed a toy experiment where we sampled from 2D Gaussians. Systematic label noise was induced by allowing label switching within a pre-defined half-space where 100% systematic label noise translates to 35% of switched labels overall (Fig. 1). The experiment was repeated 50 times using LDA as classifier (see Figure 2). We report the results in terms of AUC (ROC) when tested on the true labels (left), the matching of the new labels on the data in terms of KTA scores (center) and the percentage of true labels inferred (right). Our LATENTSVDD is less affected by variations in label noise. Other than SVDD, which infers a model of normality for each class respectively, labels are inferred for data points belonging to the same latent variable. The results show that it behaves highly accurate and much more stable when compared to the SVDD. KTA scores proof valuable for measuring label noise. However, it acts as an indicator of ground truth, not a replacement: it can increase, if the fraction of the malicious labels is higher than that of trustworthy ones.

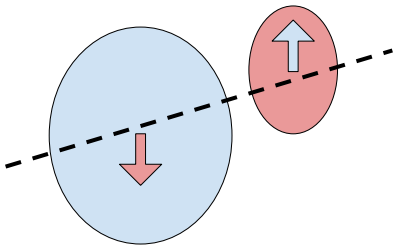

Fig. 1. Toy setting: two unbalanced classes are sampled from Gaussians, systematic label noise is induced by allowing label flips within some half-space (black dashed line, half-space for each class is indicated by the arrows)

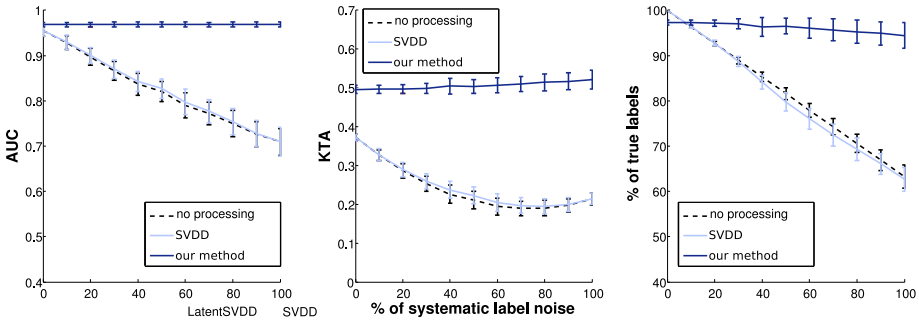

Fig. 2. Accuracy in terms of AUC tested against true labels (left). Kernel target alignment scores for the de-noised labels (center) and fraction of correctly inferred labels given the true labels (right).

3 EEG Experiment

3.1 Paradigm and Methods

Participants (N=20) were presented with a d2 test [2], a common test of visual selective attention (300 trials). Participants were asked to respond by button press as fast as possible, using their right vs. left hand for the target vs. non-target stimuli (20% vs. 80%

of trials). Feedback on speed and correctness was given 500 ms post response. Brain activity was recorded with multichannel EEG amplifiers with 119 Ag/AgCl electrodes placed according to an extended international 10-10 system, sampled at 1000 Hz and band-pass filtered between 0.05 Hz and 200 Hz.

The EEG data was divided into epochs of [-200, 500 ms] relative to the response, using the pre-response interval for baseline correction. Thus, we examined the neural data *after* the behavioral response, but *before* feedback was given. As features, we calculated the mean of the ERP signal within four neurophysiologically plausible intervals for each electrode and trial (0–80, 80–160, 200–350, 350–500 ms). In order to test class separability, we classified the EEG data using shrinkage LDA, sampling 30 times from the data set and dividing the data set into 75% training data and 25% test data. Classification was run using (a) behavioral labels, (b) the labels suggested by LATENTSVDD and (c) labels that were randomly switched with 50% probability.

3.2 Results

Classification shows that LATENTSVDD divides the neural data in a way that renders the classes clearly more distinct from each other compared to behavioral labels, reflected in higher AUC values for all but three participants (0.86 vs 0.72; red vs. blue bars in Figure 3). This is accompanied by substantially higher KTA scores for all but four participants (0.39 vs 0.01; see Figure 4), i.e. a better matching between labels and neural data. In contrast, this is not the case if labels are switched randomly: AUC values drop noticably compared to behavioral labels (0.48 vs 0.72), while KTA scores stay in the same low range.

We found that the labels retrieved by LATENTSVDD are also neurophysiologically sound. Each plot in Figure 5 shows the same data (time course at electrode Cz, participant 5), yet grouped in different classes. Classes seem relatively similar if divided into correct (green) and incorrect responses (red), based on behavioral data (Figure 5(a)). In contrast, the labels retrieved by LATENTSVDD reveal clear differences, with an error positivity P_e (red) that is much more pronounced than before (Figure 5 (d)). The inner workings of LATENTSVDD are visualized in Figure 5(b): First, the method assigns each trial to a latent variable / brain state (Figure 5(b), left). Second, LATENTSVDD uses the latent variable to assign *neural* labels (Figure 5(d), right). Red and green indicate labels that are retained by the method (brain and behavior agree); orange and light green signify trials where the labels were switched (orange to red, light green to green), which makes sense intuitively. This is confirmed when examining the differences between the two classes before and after LATENTSVDD (Figure 6). Initially (a), classes are best separated by activity in the frontal electrodes, likely related to artifacts. After LATENTSVDD (b), the most discriminative feature is an error positivity P_e (200–500ms). While the latent states are highly subject-specific, we find similar, neurophysiologcally plausible results for 16 out of 20 participants.

4 Discussion

In this paper, we proposed a measure for label noise based on KTA scores and a novel learning approach called LATENTSVDD, that allows to detect anomalies and model latent variables, which can be used to reveal latent brain states. We consider it a premier choice if labels are sparse, absent or systematically unreliable. We demonstrated

its effectiveness on EEG data recorded during a test of visual attention. The classes suggested by LATENTSVDD lead to better label-data matching and a higher separability of the data. The approach allows for a better and more meaningful experimental evaluation, not only of the neural, but also of the behavioral data: the *neural* error rate revealed by LATENTSVDD is much higher than the behavioral error rate (46.5% vs 18.05%), indicating that the brains of the participants had processed errors more often than they actually happened. Insights such as this may allow a novel view of seemingly well-known psychological phenomena.

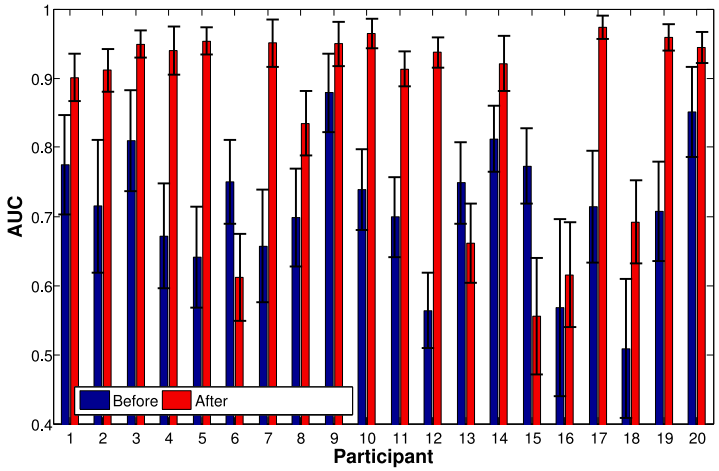

Fig. 3. Separability of the two classes by classification (AUC values), before and after running LATENTSVDD in blue and red, respectively.

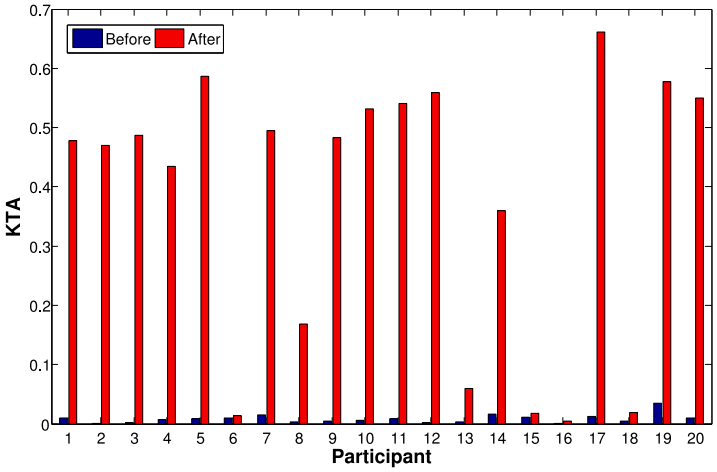

Fig. 4. Label-data matching as measured by KTA scores, before and after running LATENTSVDD in blue and red, respectively.

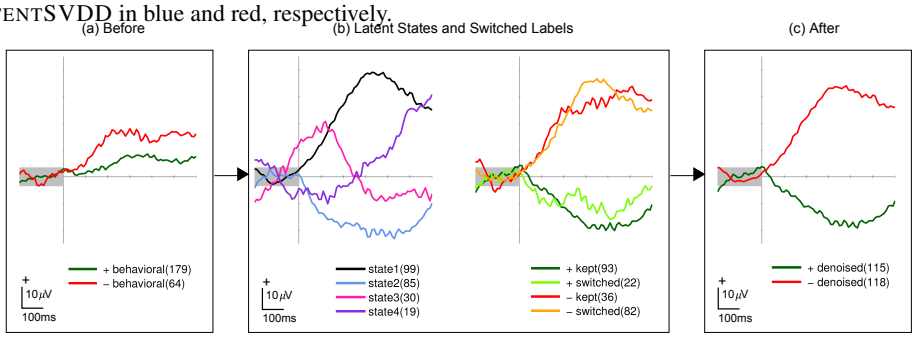

Fig. 5. Time course at electrode position Cz, [-200 600 ms] relative to the response (participant 5), with trials grouped in different classes (number of trials/class in brackets): (a) before LATENTSVDD (behavioral labels), (b) changes by LATENTSVDD (left: latent variables, right: re-assignment of labels), (c) after LATENTSVDD (denoised labels).

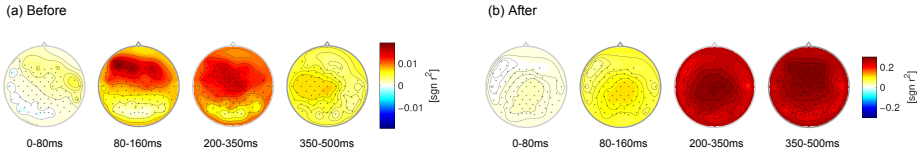

Fig. 6. Differences between classes before (a) and after (b) applying LATENTSVDD, as measured by the signed squared biserial correlation coefficient sgn-r^2 (participant 5, top view on the head with nose pointing upwards).

References

1. M. L. Braun, J. Buhmann, and K.-R. Müller. On relevant dimensions in kernel feature spaces. *Journal of Machine Learning Research*, 9:1875–1908, Aug 2008.

2. R. Brickenkamp and E. Zillmer. *D2 Test of Attention*. Hogrefe & Huber, Göttingen, Germany, 1998.
3. N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola. On kernel target alignment. In *Advances in Neural Information Proccessing Systems (NIPS)*, volume 14, pages 367–737, 2001.
4. M. Falkenstein, J. Hohnsbein, J. Hoormann, and L. Blanke. Effects of errors in choice reaction tasks on the ERP under focused and divided attention. In C. Brunia, A. Gaillard, and A. Kok, editors, *Psychophysiological Brain Research*, pages 192–195. Tilburg University Press, Tilburg, 1990.
5. M. Falkenstein, J. Hoormann, S. Christ, and J. Hohnsbein. ERP components on reaction errors and their functional significance: a tutorial. *Biol Psychol*, 51(2-3):87–107, 2000.
6. W. Gehring, M. Coles, D. Meyer, and E. Donchin. The error-related negativity: an event-related brain potential accompanying errors. *Psychophysiology*, 27:S34, 1990.
7. W. Gehring, B. Goss, M. Coles, D. Meyer, and E. Donchin. A neural system for error detection and compensation. *Psychological Science*, 4:385?–390, 1993.
8. N. Goernitz, M. Kloft, K. Rieck, and U. Brefeld. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research (JAIR)*, 46:235–262, 2013.
9. J. Hohnsbein, M. Falkenstein, and J. Hoormann. Error processing in visual and auditory choice reaction tasks. *Journal of Psychophysiology*, 3:320, 1998.
10. K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001.
11. W. Polonik. Measuring mass concentration and estimating density contour clusters – an excess mass approach. *Annals of Statistics*, 23:855–881, 1995.
12. A. K. Porbadnigk, N. Görnitz, M. Kloft, and K.-R. Müller. Decoding brain states by supervising unsupervised learning. *Journal of Computing Science and Engineering*, 2013.
13. D. M. Tax and R. P. Duin. Support vector data description. *Machine Learning*, 54:45–66, 2004.
14. I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large Margin Methods for Structured and Interdependent Output Variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.
15. V. Vapnik. *The nature of statistical learning theory*. Springer Verlag, New York, 1995.