When Brain and Behavior Disagree

Tackling systematic label noise in EEG data with Machine Learning

Anne K. Porbadnigk[1,*], Nico Görnitz[1,*],
Claudia Sannelli[1], Alexander Binder[1], Mikio Braun[1]
[1] Machine Learning Group,
Berlin Institute of Technology (TU Berlin),
Berlin, Germany
anne.k.porbadnigk@tu-berlin.de
nico.goernitz@tu-berlin.de

[*] authors contributed equally

Marius Kloft[2]
[2] Courant Institute of Mathematical Sciences and
Memorial Sloan-Kettering Cancer Center,
New York City NY, USA

Klaus-Robert Müller[1,3]
[3] Department of Brain and Cognitive Engineering,
Korea University, Seoul, Korea
klaus -robert.mueller@tu-berlin.de

***Abstract*—Conventionally, neuroscientific data is analyzed based on the behavioral response of the participant. This approach assumes that behavioral errors of participants are in line with the neural processing. However, this may not be the case, in particular in experiments with time pressure or studies investigating the threshold of perception. In these cases, the error distribution deviates from uniformity due to the heteroscedastic nature of the underlying experimental set-up. This problem of systematic and structured (non-uniform) label noise is ignored when analysis are based on behavioral data, as is being done typically. Thus, we run the risk to arrive at wrong conclusions in our analysis. This paper proposes a remedy to handle this crucial problem: 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.**

Keywords-EEG; Label Noise; Machine Learning; Unsupervised Learning; Applied Cognitive Neuroscience

I. INTRODUCTION

In recent years, there has been an increased interest in using brain-computer interfaces based on electroencephalography (EEG-BCI, e.g. [1) for novel applications, such as mental state decoding [2, 3]. In EEG experiments, each trial is associated with a stimulus/response, i.e. the stimulus presented to the participant and the behavioral response of the participant to it (for instance, in form of a button press). Typically, this information is used as a category or label and the neural data is then analyzed accordingly, in categories such as 'correct response' vs. 'incorrect response'. While this 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, cf [4] for instance) and experiments with time pressure, resulting in responses that are unreliable or even close to random guessing. A significant increase in mislabeled trials can also be caused or exacerbated when participants become distracted, bored, or sleepy (see also [5]). We assume this label noise to be systematically structured. This challenges most of the learning algorithm employed today: they struggle not only with non-uniform label noise [6, 7], but also with highly misbalanced classes (e.g., more false than correct responses in complex tasks), and the presence of brain states that are not accounted for in the experimental protocol (e.g., 'participant not on task').

In this paper, we propose an unsupervised learning algorithm called Latent Variable Support Vector Data Description (LatentSVDD) as a remedy for this challenge. LatentSVDD generealizes SVDD [8], which itself is an unsupervised anomaly detection method. The principle idea is to introduce latent variables into the SVDD. In the EEG context, these latent states can be interpreted as different brain states. In this paper, we show the usefulness of our novel framework on EEG data from a d2 attention test. In this experimental scenario, our goal is to determine whether a participant has processed a potential error on a *neural* level, which may or may not be in line with the *behavioral* level, i.e. whether the response was de facto erroneous. Neurophysiologically speaking, response errors have been found to elicite two components in the event-related potentials (ERPs): the error negativity and the error positivity [9,10]. While the former has been attributed to the comparison process rather than its outcome, the latter has been suggested to be related to error or post-error processing [11]. Therefore, we concentrate on the error positivity in the following, i.e. a centro-parietal maximum that has been found to occur 200-500ms post response.

II. LEARNING METHODOLOGY

In the following, we consider a learning scenario that is characterized by labels that have varying levels of reliability. As a remedy, we propose a measure based on kernel target alignment scores (KTA) and a novel, 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.

A. *Kernel Target Alignment (KTA)*

We are given N labels $y \in \{+1,-1\}^N$ and a Gram matrix $K \in M(NxN,R)$. Kernel target alignment (KTA) [12] is a method to measure the fit between the gram matrix and the label set. It is defined as $KTA(K,y) = \frac{< K, yy^T >_F}{\|K\|_F \|yy^T\|_F}$

A high value is achieved, if data points of one class lie nearby and data points of opposite classes are far away. 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.

B. *Latent Variable Support Vector Data Description (LatentSVDD)*

Our approach is based on the paradigms of support vector learning [13,14], density level set estimation, support vector data description (SVDD) [8,15] and extensions [16]. We are given N data points $x_1,...,x_N$, where x_i lie in some input space R^d. The data is usually mapped from the input space into some feature space $\phi : R^d \to F_c$. In SVDD, the goal is to find a center c and a radius R of a hypersphere, that contains the bulk of the data: $f : R^d \to R, x \mapsto \|c - \phi(x)\|^2$. Thus, the optimization problem can be stated as

Minimize R^2

Such that $\|c - \phi(x_i)\|^2 \le R^2$

In this paper, we extend the classical mapping of SVDD by including a latent variable $z \in Z$ and a joint feature map $\Psi : R^d xZ \to F$. Consequently, the resulting model becomes more expressive:

$$f : R^d \to R, x \mapsto \min_{z \in Z} \|c - \Psi(x,z)\|^2 .$$

Here, we define our joint feature map as a variant of the multi-class joint feature map $\Psi(x,z) = \phi(x) \otimes \delta(z_k, z)$ with $k \in \{1,...,12\}$ (i.e. 12 latent brain states, which is more than we expect and thus serves as an upper bound). We train our method on all available data points, which results in anomaly scores and latent variables for each of them. Labels are assigned depending on maximum KTA scores. Figure 1 visualizes the main concepts.

In order to gain insights on the behavior of the proposed model, we generated toy data and applied the LatentSVDD. The outcome can be seen in Figure 2. As depicted on top, the level of anomaly (visualized by dot size and color) is reflected well by the scores assigned by LatentSVDD. The plot at the bottom encodes the separation of the data points by latent states.

Supported by the German Bundesministerium für Bildung und Forschung (BMBF FKZ 01GQ0850, 01IB001A, 01IB10003B), the German Science Foundation (DFG MU 987/6-1, RA 1894/1-1), and the World Class University Program of the Korea Research Foundation (R31-10008)

Figure 1. Main idea: we infer a model of normality by learning a hypershere containing most of the data.

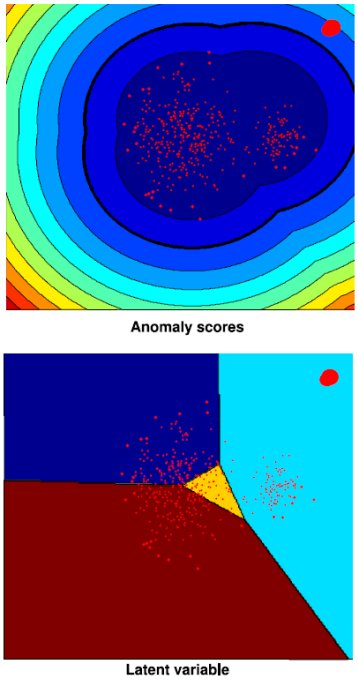

Figure 2. Model output on toy data. Top: anomaly scores (dot size and color corresponds to degree of anomaly). Bottom: areas of active latent variables.

III. EEG EXPERIMENT

Participants (N=20) were presented with a d2 test [17], 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 500ms 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.

We examined the neural response that was elicited by receiving feedback. For this, the EEG data was divided into epochs of [-200, 500ms] relative to the onset of feedback presentation. These epochs were then baseline corrected, using the 200ms interval prior to feedback. Artifact rejection was performed using a min-max criterion and a variance criterion (trials and channels). In order to reduce dimensionality [18], we calculated 9 features per epoch, which were used as input both for LatentSVDD and classification. For this purpose, the interval [0 500ms] was divided in a total of 10 non-overlapping intervals, each with a length of 50ms. We then calculated the

mean signal in each of these intervals and subsequently, the gradient between these means.

In order to test the separability of classes, 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, and (b) the labels inferred by LatentSVDD. Our goal was to apply LatentSVDD in order to divide the trials according to whether an error was processed on a neural level or not.

IV. Results

A. *Class Re-assignment and Anomalous Trials*

On average, LatentSVDD flipped the labels for 35.94% of all trials. This resulted in a neural error rate of 31.18%, compared the lower behavioral error rate (18.05%). Based on the anomaly score that LatentSVDD returns for each trial, we rejected a small percentage of trials for each participant. For the majority of participants, there are only few trials with high anomaly scores, with a steep drop-off compared to the remainder of the trials (cf. Figure 3). Visual inspection revealed that the results are plausible from a neuroscientific point of view: the rejected trials show typical artifacts (eye blinks, voltage drifts with respect to all electrodes or a single electrode) that have escaped the conventional artifact rejection run prior to applying LatentSVDD.

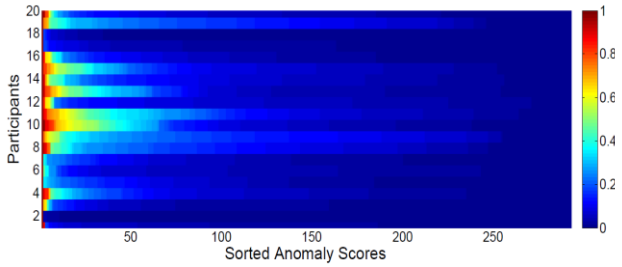

Figure 3. Sorted anomaly scores for each data point of each participant.

B. *Quantitative Assessment*

We quantified the benefits of LatentSVDD using KTA scores and linear classification (LDA). Both measures confirm that the labels assigned by LatentSVDD allow a much better separation of the data than behavioral labels for all 20 participants. As can be seen in Figure 4.A, LatentSVDD renders the classes clearly more distinct from each other, reflected in higher AUC values (0.95 vs. 0.60). This is accompanied by substantially higher KTA score for all participants, as can be seen in Figure 4.B.

C. *Neurophysiological Assessment*

While AUC and KTA scores help quantify the positive effect of LatentSVDD, we found the results also to be neurophysiologically sound. In the following, we discuss this for our methodology at the example of participant 4. The different steps of our methodology are visualized in Figure 5.

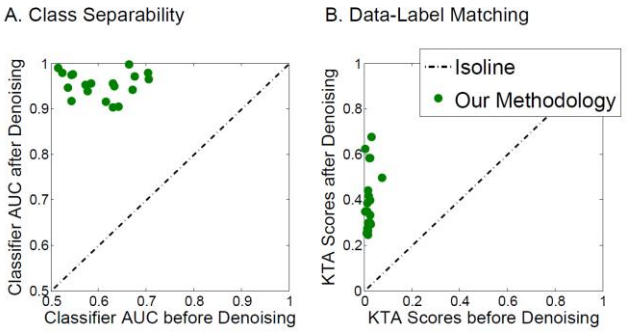

Figure 4. A. Separability of the two classes by classification (AUC values), B. Label-data matching as measured by KTA scores, as measured before and after running LatentSVDD (x-axis vs. y-axis).

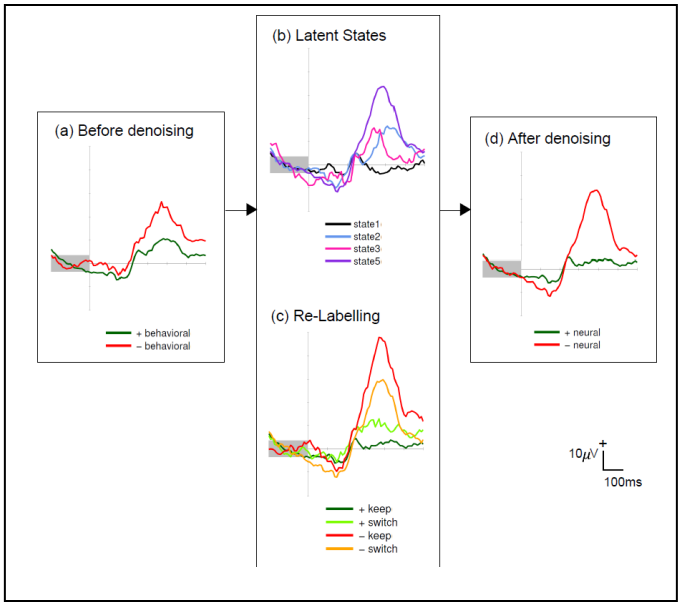

Figure 5. Time course at electrode position Cz, [-200, 600ms] relative to the response (participant 4), with trials grouped in different classes: (a) before LatentSVDD (behavioral labels; green/red: correct/incorrect response), (b) latent states revealed by LatentSVDD, (c) suggested re-assignment of labels, (d) after LatentSVDD (denoised labels).

Each plot shows the same data (time course at electrode Cz, participant 4), 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 (red) that is much more pronounced than before (Figure 5(d)). The inner workings of LatentSVDD are visualized in the middle of Figure 5: First, the method assigns each trial to a latent variable / brain state, as can be seen in Figure 5(b). The state with the highest amplitude (purple) corresponds to typical error processing, with a clear positive component. A clear positivity also occurs in two other states (blue and pink), yet less pronounced and with different latencies. In contrast, no error has been processed in the fourth state (black). Based on the latent variable, a subset of trials is

then re-assigned (Figure 5(c)). Red and green indicate labels that are retained, orange and light green signify trials where the labels were switched (orange to red, light green to green). As can be seen, the re-assignment makes sense intuitively. Finally, Figure 5(d) shows the denoised data, which reveals a more pronounced error positivity (red) than before. While the latent states themselves are highly subject-specific, we find similar, neurophysiologically plausible results for the majority of participants.

V. Discussion

In this paper, we propose a measure for label noise, using KTA scores as well as a novel learning approach called LatentSVDD. The latter allows us 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. In this paper, we demonstrate its effectiveness on an EEG data set, recorded during a test of visual attention. We show that the classes inferred by LatentSVDD lead to better label-data matching and a substantially higher separability of the data (assessed with linear classification; rise in the mean AUC from 0.60 to 0.95). Visual inspection of the data shows that the class assignments by our method are neurophysiologically plausible, leading to more easily interpretable brain states that may subsequently allow for a better and more meaningful experimental evaluation. Interestingly, the *neural* error rate revealed by LatentSVDD is much higher than the behavioral error rate (31.18% vs 18.05%), indicating that the brains of the participants had processed errors more often than they actually happened. Thus, our approach allows for a better and more meaningful experimental evaluation, not only of the neural, but also of the behavioral data.

Acknowledgments

This work was supported by the German Bundesministerium für Bildung und Forschung (BMBF FKZ 01GQ0850, 01IB001A and 01IB10003B), the German Science Foundation (DFG MU 987/6-1, RA 1894/1-1), and the World Class University Program of the Korea Research Foundation (R31-10008). Marius Kloft acknowledges a postdoctoral fellowship by the German Research Foundation (DFG).

References

[1] Dornhege G, Millán J del R, Hinterberger T, McFarland D, Müller KR, Eds., Toward Brain-Computer Interfacing. Cambridge, MA: MIT Press, (2007)

[2] Blankertz B, Tangermann M, Vidaurre C, Fazli S, Sannelli C, Haufe S, Maeder C, Ramsey LE, Sturm I, Curio G, Müller KR, The Berlin Brain-Computer Interface: Non-Medical Uses of BCI Technology, Front Neuroscience, 4:198, (2010)

[3] Müller KR, Tangermann M, Dornhege G, Krauledat M, Curio G, Blankertz B, Machine learning for real-time single-trial EEG analysis: From brain-computer interfacing to mental state monitoring, J Neurosci Meth 167, 82-90 (2008)

[4] Porbadnigk AK, Treder MS, Blankertz B, Antons JN, Schleicher R, Möller S, Curio G, Müller KR, Single-trial analysis of the neural correlates of speech quality perception, J of Neural Engineering, 10(5), 056003 (2013)

[5] Porbadnigk AK and Görnitz N, Kloft M, Müller KR, Decoding brain states by supervising unsupervised learning, J of Computing Science and Engineering, 7(2), 112-121 (2013)

[6] Braun ML, Buhmann J, Müller KR, On relevant dimensions in kernel feature spaces. Journal of Machine Learning Research, 9:1875–1908 (2008).

[7] Montavon G, Braun ML, Krueger T, Müller KR, Analyzing Local Structure in Kernel-based Learning: Explanation, Complexity and Reliability Assessment. Signal Processing Magazine, IEEE, 30(4):62-74 (2013).

[8] Tax DM, Duin RP, Support vector data description. Machine Learning, 54:45–66 (2004).

[9] Gehring W, Coles M, Meyer D, Donchin E, The error-related negativity: an eventrelated brain potential accompanying errors. Psychophysiology, 27:S34 (1990).

[10] Gehring W, Goss B, Coles M, Meyer D, Donchin E, A neural system for error detection and compensation. Psychological Science, 4:385–390, (1993).

[11] Falkenstein M, Hoormann J, Christ S, Hohnsbein J, ERP components on reaction errors and their functional significance: a tutorial. Biol Psychol, 51(2-3):87–107 (2000).

[12] Cristianini N, Shawe-Taylor J, Elisseeff A, Kandola JS. On kernel target alignment. In Advances in Neural Information Proccessing Systems (NIPS), 14:367–737 (2001).

[13] Vapnik V, The nature of statistical learning theory. Springer Verlag, New York (1995).

[14] Müller KR, Mika S, Rätsch G, Tsuda K, Schölkopf B, An introduction to kernel-based learning algorithms. IEEE Transactions on Neural Networks, 12(2):181–201 (2001).

[15] Polonik W, Measuring mass concentration and estimating density contour clusters – an excess mass approach. Annals of Statistics, 23:855–881 (1995).

[16] Görnitz N, Kloft M, Rieck R, Brefeld U, Toward supervised anomaly detection. Journal of Artificial Intelligence Research (JAIR), 46:235–262 (2013).

[17] Brickenkamp R, Zillmer E, D2 Test of Attention. Hogrefe&Huber, Göttingen, Germany (1998).

[18] Blankertz B, Lemm S, Treder MS, Haufe S, Müller KR, Single-trial analysis and classiffication of ERP components - a tutorial, Neuroimage, 56, 814-825 (2011)