Martin Patrick Pauli*, Constantin Pohl, Martin Golz

Balanced Leave-One-Subject-Out Cross-Validation for Microsleep Classification

Abstract: Inter-individual differences in the feature distribution of electroencephalograms (EEG) during microsleep (MS) raise questions about the generalizability of the methodology, because methodology should have the same validity for subjects to be included in the analysis in the future as for subjects included so far. We address this question using leave-one-subject-out cross-validation (LOSO CV) to simulate inclusion of test data from future subjects. Investigations are based on EEG of 70 subjects across four studies conducted in our driving simulation lab. 9,297 MS and 10,264 counter-examples of sustained attention (SA) were included in the analysis. Three variants of sample balancing are compared: (1) none, (2) overall balancing, and (3) between-subjects balancing. Resulting classification accuracies range from 80% to 100% for 67 subjects, but for 3 subjects they are at 66.5%, 73.1%, 78.3%. Averaged across subjects, accuracies were (1) 90.2%, (2) 90.2% and (3) 89.1%. Thus, balancing might not be essential if LOSO CV is performed over a sufficiently large number of subjects. Results of standard CV must be regarded as optimistically biased.

Keywords: Electroencephalogram, microsleep, machine learning, cross validation, gradient boosting machine.

<https://doi.org/10.1515/cdbme-2021-2038>

1 Introduction

In biosignal processing, empirical validation plays a key role as an objective method in addition to subjective and heuristic validation. Its basic assumption is that the sample is representative and that analysis methods and their parameterisation have been chosen optimally. Meeting these assumptions cannot usually be guaranteed. In the field of computational intelligence, cross-validation (CV) is established as de facto standard [1]. Data are split into training and test sets to adapt internal hypothesis parameters during training and to estimate predictive performance of the hypothesis during testing. The underlying assumption is that samples are drawn independently and from an identical distribution. However, this cannot be

guaranteed as hierarchical structures, spatio-temporal correlations or non-stationary processes may be present.

Because of hierarchy argument, it has been recommended to use leave-one-subject-out CV (LOSO CV) instead of standard CV for multi-subject clinical diagnostic studies [2]. Training data consists of all but one subject's data. This division is repeated after training and validation until each subject's data once formed the test set. With LOSO, sample bias can be estimated because the absence of a subject is simulated in the analysis. Moreover, it is possible to assess whether data of all subjects are from the same distribution. In models with identity-confounding effects, however, LOSO fails [3].

Unequal a priori probabilities also have biasing influence [1]. Thus, class-related data sizes should be balanced. This is investigated by using three variants of balancing the training sets; test sets were always balanced. In LOSO 1, there was no balancing. In LOSO 2, balancing was done in total across all subjects; each subject has a different number of examples. In LOSO 3, each subject has equal number of examples for each class. This leads to a dramatic loss of data, as the subject with least number of MS examples determines the size of the training set. Balancing is always performed by deleting randomly selected examples of the majority class until the size of the minority class is reached. This contribution aims at investigating the influence of balancing the training set on the results of LOSO CV. To our knowledge, such investigations have not yet been carried out for microsleep, but in other EEG studies, among others for epilepsy diagnostics [4], where LOSO as well as balancing was also taken into account.

2 Material

For data analyses, EEG from four driving simulation studies (A to D) from 2007, 2008, 2016, 2018 with a total of 79 participants was used. In A and B, recording device was SIGMA PL Pro® (SIGMA Medizintechnik GmbH, Germany) and in C and D, SOMNOscreen® (SOMNOmedics GmbH, Germany)

*Corresponding author: **Martin Pauli:** University of Applied Sciences, Schmalkalden, Germany, mp.pauli@hs-sm.de

2nd Author Constantin Pohl, 3rd Author Martin Golz: University of Applied Sciences, Schmalkalden, Germany

Figure 1: Absolute and relative numbers of MS and SA examples and threshold value (black line) versus subject ID.

each with nine electrodes. Analyses were limited to the following six electrodes: Fp1/2, C3/4, O1/2. Moreover, three videos of the driving scene, head-shoulder and eye region were recorded, as well as driving simulator data. They were used by an expert to visually assess driver behavioural states: microsleep (MS), sustained attention (SA) and not clearly evidenced behaviour. Strict criteria have been applied, so that not clearly evidenced states were selected more often than necessary, especially in cases of high fatigue. They were excluded from analyses. A total of 9,297 MS and 10,264 SA were selected.

The three LOSO variants mentioned above resulted in the following data sizes: (MS / SA): (1) 9,297 / 10,264; (2) 9,297 / 9,297; (3) 2,450 / 2,450.

Subjects were young adults (30 ♀, 49 ♂, 23.9 ± 2.8 years). They were monitored by actometers so that there was an assured minimum time since sleep (A, B: 14 h; C, D: 16 h) before starting first driving session (A, B: 10:30 pm; C, D: 1:00 am). Hourly driving sessions followed (A, B: 8; C, D: 7), each lasting 40 min, so that another important factor of fatigue, time on task, was relatively high (A, B: 320 min; C, D: 280 min). The circadian trough between 4 and 6 am was traversed in all studies. These factors led to a high fatigue and thus to a high number of MS, which is of great advantage for computational intelligence (CI) methods.

3 Methods

Segments 4 s long were taken from the EEG, with an offset of 1.5 s between segment onset and MS onset, so that it was about in the mid-segment. Both parameters were determined by empirical optimisation. For SA counterexamples, segments of the same length that were entirely within time intervals where attentive driving could be visually observed were used. After visual elimination of artefact segments, trend was eliminated and Hamming taper applied. Discrete Fourier transform was then used to estimate the power spectral densities and averaged after logarithmic scaling in 1 Hz wide frequency intervals ranging from 0 to 30 Hz. Thus, 30 LogPSD values were estimated in each of the six EEG channels; they form a feature vector $\mathbf{x}_n$ of 180 components, which serve as input vector for CI methods. The target variable is binary: $t_n = +1$ or $t_n = -1$ for the behavioural state MS or SA, respectively. This led to the supervised learning dataset:

$$S = \{(\mathbf{x}_n, t_n) | \mathbf{x}_n \in \mathbb{R}^d, t_n \in \{-1, +1\}, n \in [N]\} \quad (1)$$

with $d = 180$, $N = 19,030$, and $[N]$ being the set of data indices $\{1, 2, \dots, N\}$.

The CI method used was mainly LightGBM [5]; Optimised Learning Vector Quantisation (OLVQ1) and Support Vector Machines (SVM) were also used for comparison. LightGBM belongs to the class of ensemble methods in which

γ -weak learning algorithms (γ -WLA) are combined linearly. In each boosting iteration (index i) the ensemble increases by one γ -WLA representing hypothesis h_i and having empirical loss E_i only slightly smaller than that of random choice:

$$E_i = \sum_{n=1}^N v_n(i) \mathbb{I}_{[h_i(x_n) \neq t_n]} \leq \frac{1}{2} - \gamma \quad (2)$$

with indicator function $\mathbb{I}_{[\cdot]}$ and data weights $v_n(i)$. In addition, risk-dependent hypothesis weights w_i will be used such that a small E_i gives rise to high weights and vice versa:

$$w_i = \frac{1}{2} \ln \left(\frac{1}{E_i} - 1 \right) \quad (3)$$

An overall ensemble hypothesis is generated by linear combination, where $\text{sgn}(\cdot)$ is the sign function and $\mathbf{x}$ is any input vector for which an output is sought:

$$h_S(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^{N_i} w_i h_i(\mathbf{x}) \right) \quad (4)$$

Index S emphasises that the overall hypothesis was obtained empirically on basis of the sample S (eq 1). At the end of each iteration, data weights are adjusted by giving each example $(\mathbf{x}_n, t_n) \in S$ for which h_i was incorrect ($h_i(\mathbf{x}_n) \neq t_n$) a higher data weight $v_n(i)$ and each correctly classified ($h_i(\mathbf{x}_n) = t_n$) examples gets a lower data weight:

$$v_n(i+1) = \frac{v_n(i) e^{-w_i t_n h_i(\mathbf{x}_n)}}{\sum_{j=1}^i v_n(j) e^{-w_j t_n h_j(\mathbf{x}_n)}} \quad \forall n \in [N] \quad (5)$$

It has been proven that the empirical mean loss L_S has an upper bound that decreases exponentially with increasing number of iterations N_i and thus of hypotheses:

$$L_S(h_S) = \frac{1}{N} \sum_{n=1}^N \mathbb{I}_{[h_S(\mathbf{x}_n) \neq t_n]} \leq e^{-2\gamma^2 N_i} \quad (6)$$

This way, an ensemble of weighted γ -WLA represents a strong learning algorithm. Decision trees [6] have been used here as γ -WLA. However, their efficiency is limited by number of features d and by data set size $N = |S|$, so it is in the order of magnitude of $\mathcal{O}(Nd)$ for a decision (split), because a histogram must be created for each feature over all examples and the information gain (split criterion) must then be estimated. This can lead to considerable computational demands for large data sets. The gradient boosting method (GBM) [6] uses gradient descent and exploits the fact that for quadratic objective functions $J = \frac{1}{2} \sum_{n=1}^N (t_n - h_i(\mathbf{x}_n))^2$ the first derivative with respect to h_i is equal to the residual of target value t_n and hypothesis output $h_i(\mathbf{x}_n)$. For hypothesis-internal parameters θ_j , linear approximation yields the following adaptation rule:

$$\theta_{i,j+1} = \theta_{i,j} - \eta \frac{\partial J}{\partial \theta_{i,j}} \quad (7)$$

The choice of step size η ($0 < \eta < 1$) is important for learning success. If η is chosen too small, the objective function is only slightly reduced per iteration, which leads to inefficient training especially in plateau regions. If η is chosen too large, GBM convergence deteriorates. If decision trees are used as γ -WLA, then the parameters θ_j are the threshold values of the individual decisions (splits) for each feature. LightGBM is a computationally light GBM algorithm [5] using two concepts:

1. Gradient-based one-side sampling: significant amounts of low-gradient data examples are removed from computation of the information gain.
2. Exclusive feature bundling: pairs of mutually exclusive features are removed. In such pairs, non-zero values very rarely occur simultaneously.

Both concepts reduce effectively necessary sizes of N and d . Thus, they reduce training time by up to a factor of 20 [5].

4 Results

The numbers of examples for both classes (MS, SA) vary greatly from subject to subject (Fig. 1). However, sufficiently high numbers are important in order to have for each subject a more or less representative sample for the CI methods. For this reason, subjects whose absolute MS numbers were below a freely chosen threshold were excluded from further analyses; this was the case for nine subjects. The relative distribution of the numbers of both classes, the so-called balancing, was also used as an exclusion criterion, again with a freely chosen threshold. All subjects who did not meet the criterion were already excluded due to the first criterion. As a result, 70 subjects were included in further analyses.

In order to assess the generalisation ability of LightGBM, mean classification accuracies over the test set were considered. The mean was computed over five training repetitions to keep low the influence of the sample order on the results. For visual assessment, subject ID was reordered w.r.t. level of accuracy achieved leading to a monotonically increasing graph (Fig. 2). The accuracies over the significantly larger training sets characterise only the adaptivity of LightGBM, which is very high and therefore LightGBM has a disadvantageous tendency to overfitting rather than underfitting. Training accuracies were always between 99.6% and 100% and are not shown here. The mean of all results (Fig. 2, lower red line) and, due to the outliers, the median (upper red line) were calculated. For comparison, results of standard cross-validation, i.e. repeated

Figure 2: Individual mean classification accuracy of test sets versus subject ID for three LOSO methods. ID was reordered to have monotonically increasing accuracy. Mean and median across all subjects (lower/upper red line) and Non-LOSO result (cyan line) are shown for comparison.

random subsampling, are shown (cyan line). In the latter case, data from all test subjects are included in the training set, so that there is no longer a subject-separated analysis here. Results of LOSO 1 to 3 do not differ significantly, except for approx. 25 subjects with mean accuracies below 90 %.

Mean values for LOSO 1 to 3 are approx. 2 % lower compared to those in Fig. 2 when a uniform overall hypothesis is finally calculated for all subjects. LOSO variants yielded: (1) 90.2%, (2) 90.2% and (3) 89.1%.

Two other CI methods were used in standard cross-validation to provide comparative evidence on whether LightGBM is sufficiently accurate. Mean accuracies are: OLVQ1 86.4 %, SVM 94.16 %, LightGBM 94.14%.

5 Conclusions

In event-oriented psychophysiological studies, the problem of large scatter in the number of events arises. On one hand, one is forced to include as many examples as possible in data analysis so that CI methods can unfold their potential, and on other hand, especially in subject-separated validation analysis, subjects should not be included whose event numbers are so small

that they can no longer be assumed as representative. Such a trade-off is proposed here; LOSO analyses based on this show that there are no significant differences between balancing variants. Even if identical balancing is carried out for each subject and across subjects, thus accepting a large reduction in data size, mean accuracy is almost equally high. This indicates that EEG during MS is characteristically different from SA and can be learned in a generalizable and subject-independent manner.

Classification accuracy was investigated; analysis of sensitivity, specificity and receiver operating characteristics could have further increased the complexity of the results to be assessed and shall be reserved for future investigations.

The most important result is that compared to an earlier LOSO analysis with less than half of subjects [7], accuracies achieved are now slightly higher and that this even applies to an extremely reduced sample (LOSO 3). Obviously, added subjects provide some helpful information to better predict the subject held out in the training. Future studies should therefore include an even larger number of subjects.

Author Statement

Research funding: The author state no funding involved. Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent has been obtained from all individuals included in this study. Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

References

- [1] Kuhn M, Johnson K. Applied Predictive Modelling. Springer, NY 2013.
- [2] Saeb S, et al. The need to approximate the use-case in clinical machine learning. *Gigascience* 2017; 6(5):1-9, gix019.
- [3] Little MA, et al. Using and understanding cross-validation strategies. *GigaScience* 2017; 6(5):1-6, gix020.
- [4] Thomas et al. Automated adult epilepsy diagnostic tool based on interictal scalp electroencephalogram characteristics: six-center study. *Int J Neural Syst* 2021; 31(5):2050074.
- [5] Ke G, et al. LightGBM: A highly efficient gradient boosting decision tree. *Adv Neural Inf Proc Syst* 2017; 30:3146-54.
- [6] Friedman J. Greedy function approximation: a gradient boosting machine. *Ann Statist* 2001; 29:1189-232
- [7] Golz M, Schenka A, Haselbeck F, Pauli MP (2019) Inter-individual variability of EEG features during microsleep events. *Curr Dir Biomed Engin* 2019; 5(1):13-6.