Proceedings on Automation in Medical Engineering
Proc AUTOMED, Vol 2, No 1 (2023): Article Ref 718
www.journals.infinite-science.de/automed/article/view/718

Infinite Science Publishing

On the impact of feature reduction on leave-one-subject-out cross-validation

M. P. Pauli [1]* and M. Golz[1]

[1] Department of Computer Science, University of Applied Sciences, Schmalkalden, Germany
** Corresponding author, email: mp.pauli@hs-sm.de*

Abstract: The high inter-individual variability of the electroencephalogram (EEG) is investigated in this contribution using leave-one-subject-out cross-validation (LOSO CV). The question of whether feature reduction can significantly increase the generalization ability of LOSO CV is addressed or whether feature reduction causes too high loss of relevant features, thus worsening the results. EEG recordings from three driving simulation studies are analyzed, in which microsleep (MS) and sustained attention (SA) were observed in 66 young drivers, some with very high fatigue. The gradient boosting machine LightGBM, was used as a classifier for discriminating MS and SA. The results show that the mean classification accuracies at validation sets have been found to be 90.8±0.8% for the standard CV and 86.4± 11.0% for the LOSO-CV. Through three different feature reduction criteria, a total of 57 different reductions were performed with different thresholds, but there were no significant improvements in the mean LOSO-CV accuracy. If the reduction was chosen too high, i.e. more than 96% of the features were not processed, then significant reductions of the mean classification accuracies down to 79.1% were obtained.

I. Introduction

Due to the high inter-individual variability of many observable variables in life sciences, it is of great importance in the case of machine learning analyses that leave-one-subject-out cross-validation (LOSO CV) is also investigated instead of standard cross-validation. This will determine the impact of each individual's data on the model fit. The reason is that LOSO CV keeps all of an individual's data completely out of the training process in order to determine, in the validation step, how accurately a model adapted on all other individuals' data works for that individual. This is essential for statistical analyses, because training and validation sets must be independent of each other. A data leakage [1] in the training process must be avoided. LOSO-CV avoids the so-called group leakage. On a larger dataset, we demonstrate that for detection of microsleep in EEG, the LOSO-CV leads to lower mean classification accuracies with significantly increased variability compared to the standard CV [2]. Thus, the optimistic bias of the standard CV must be considered critically.

II. Material

EEG recordings from three driving simulation studies with a total of 66 participants were included in the data analysis. EEG was recorded in each of the Fp1, Fp2, C3, C4, O1, and O2 channels; Cz served as the reference electrode. Eye movements were recorded using vertical and horizontal EOG; however, these are excluded from the analysis because they were needed for the independent expert evaluation of the two behavioral states of microsleep (MS) and sustained attention (SA). Video recordings of the eye region, head-shoulder region, and driving scene were also used for the expert evaluation.

Three important factors influencing fatigue were set high: (1) the time since sleep was at least 14 h; (2) the study period covered the circadian trough between 4 and 6 am; and (3) the accumulated time on task was relatively high, greater than or equal to 280 min [2]. These three factors, and especially the monotonicity experience in the driving simulator, resulted in high to extreme fatigue and consequently a high number of MS, which is important because sample size is crucial for approximately correct learnability [3].

III. Methods

The modified periodogram was estimated directly from 4-second EEG segments after trend elimination and Hamming tapering, and then the LogPSD was calculated in narrow 1 Hz spectral bands in the range from 0 to 40 Hz [4]. These 40 variables for each of the above 6 channels formed the $6 \cdot 40 = 240$ components of the feature vectors, which in turn are the input variables of machine learning.

Machine learning was performed using the gradient boosting method LightGBM, a comprehensive framework available for download from Microsoft Research Inc [5]. In particular, it was trimmed for efficiency so that the extensive LOSO computations could be performed in acceptable time using PC grid computing. LightGBM has extensive functionalities to parameterize and configure it for various use cases. Gradient boosting belongs to ensemble methods where many weak learning algorithms are trained sequentially and then their hypotheses are weighted and fused into an overall hypothesis. As weak learning algorithms, decision trees with the Gini impurity criterion were used here to automatically find the decision thresholds.

As standard cross-validation (CV) method, the K-fold CV or the RRSS CV (repeated random subsampling) is usually used. Since the latter allows a higher number of repeated training runs, it leads to statistically more reliably estimable results. With RRSS CV, in step (1) the data index is randomly permuted so that in training the data are used in randomized order. In step (2), with a quantity ratio of, e.g. 8:2 or 9:1, the data set is split into training and validation sets. During the following step (3), the LightGBM training is performed, for which only the training set is used. In step (4), each element of the training set is processed in the classifier recall (LightGBM recall) and counted how many true positive and true negative classifications occur with respect to the size of the set to finally obtain the training accuracy a_T. The same calculations are performed in step (5) for the elements of the validation set without adapting LightGBM to obtain the validation accuracy a_V. All steps are repeated M times, so that finally M different training and M different test accuracies are available. M can be chosen freely and is usually between 10 and 100; here $M = 25$ was chosen. In step (6) the arithmetic mean and the standard deviation are estimated over all M values.

In LOSO-CV, step (2) is executed first and the data of one individual is declared as validation set and the P-1 remaining data is declared as training set (P = number of individuals). Then, step (1) of permuting the data indices of the training set is performed. This is followed by steps (3) to (5). These five steps are repeated P times, so that finally P different training and P different validation accuracies are available. Finally, step (6) is performed as above, leading to the mean accuracies $\overline{a_T}, \overline{a_V}$ for training and validation sets, respectively. The former can be used to estimate the adaptivity of the classifier for the given training sets, which is usually very high for LightGBM; $\overline{a_T}$ indicates whether the learning processes have been successful and whether the predefined configuration settings are successful. However, the real importance comes to $\overline{a_V}$, since it can be used to estimate the generalization ability, i.e., how successfully LightGBM can be applied to data of future individuals.

Feature reduction was performed using (a) the split criterion, i.e. the number of splits on a feature. If this number was below a threshold, the feature was considered irrelevant and eliminated. Alternatively, (b) the gain criterion, i.e., the contribution of the feature to the classification accuracy, was used. As criterion (c), the conjunctive combination of criteria (a) and (b) was used. As threshold values, related to the maximum of the criterion among all features, the median and mean were used and the relative values: 0.1%, 0.25%, 0.5%, 0.75%, 1.0%, 2.5%, 5%, 10%, 10.5%, 11%, 12%, 13%, 14%, 15%, 20%.

III. Results and discussion

The mean classification accuracy on validation sets was $\overline{a_V} = 90.8 \pm 0.8\%$ in case of standard CV (RRSS) (Fig.1, red, ID=0). In case of LOSO-CV, it was $86.4 \pm 11.0\%$ (Fig.2,ID=1). The significantly higher standard deviation is caused by a large number of individuals that are not sufficiently accurately represented by the data set of the other $P - 1$ individuals. For three individuals in which particularly low accuracies were achieved, the causes were poor signal quality [2]. The hypothesis that selective feature reduction would increase the robustness of the LightGBM models could not be confirmed. The results were sorted in Fig. 1 such that as the index increased, feature reduction was strengthened and the number of features n_F decreased. If feature reduction resulted in equally large n_F, results were nearly the same; so for clarity, they were not included in Fig.1. It can be seen that for almost all reductions there is no significant change in the mean accuracies $\overline{a_V}$ (Fig.1, ID=2,3,...,20). Only for drastic reductions, where more than 96% of the features were eliminated (Fig.1, ID=21-23), there were significant changes, but lower accuracies $\overline{a_V}$. Increased $\overline{a_V}$ are only slightly indicated when about 68% to 75% of the features were eliminated (Fig.1, ID=11-15). Here, the standard deviation was also slightly reduced.

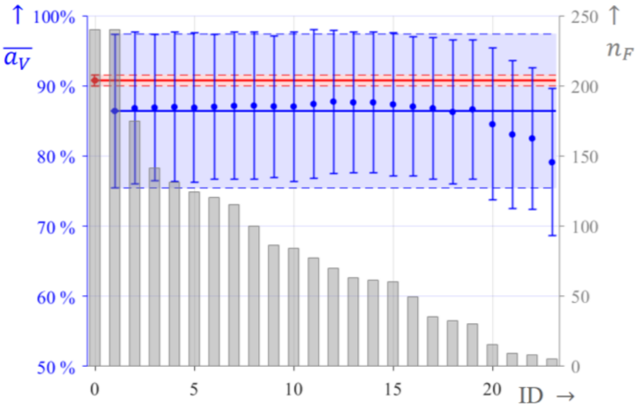

Figure 1: Mean and standard deviations of the mean validation accuracy $\overline{a_V}$ (dots, error bars) versus identification index ID of feature reduction. Results were ordered such that the number of features n_F (gray bars) decreases strictly monotonically with ID. The RRSS-CV reference (red, ID=0) and the LOSO-CV reference (blue, ID=1) were performed without feature reduction (n_F=240). The LOSO-CV results $\overline{a_V}$ with feature reductions (blue, ID=2-23) are similar to $\overline{a_V}$ of the reference (ID=1).

IV. Conclusions

Feature reduction has the potential but no guarantee to generate more robust and generalizable hypotheses. Using a relatively large data set we demonstrated that high dimensionality is not only a curse, but also a blessing.

Author's Statement

The author state no funding involved as well as no conflict of interest. Informed consent has been obtained from all individuals included in this study. 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] Kaufman S, Rosset S, Perlich C, Stitelman O (2012) *Leakage in data mining: Formulation, detection, and avoidance.* ACM Trans Knowl Discov Data (TKDD), 6 (4) pp 1-21.

[2] Pauli MP, Pohl C, Golz M (2021) *Balanced Leave-One-Subject-Out Cross-Validation for Microsleep Classification.* Curr Direct Biomed Engin **7** (2) pp 147-150.

[3] Shalev-Shwartz S, Ben-David S (2014) *Understanding machine learning: From theory to algorithms*. Cambridge Univ Press.

[4] Pauli MP, Pohl C, Golz M (2022) *Optimal EEG Segmentation for Microsleep Detection Based on Machine Learning*. Curr Direct Biomed Engin 8 (2) pp 749-752.

[5] Ke G, et al. (2017) *LightGBM: A highly efficient gradient boosting decision tree.* Adv Neural Inf Proc Syst 30 pp 3146-3154.