Kalpa Publications in Computing

Volume 14, 2023, Pages 170–173

Proceedings of V XoveTIC Conference. XoveTIC 2022

Comparison of ensemble-combination approaches in an automatic sleep staging inter-database generalization task

*

Adriana Anido-Alonso[1] and Diego Alvarez-Estevez[2]

[1] CITIC, Universidade da Coruña, A Coruña, Spain
adriana.anido@udc.es
[2] CITIC, Universidade da Coruña, A Coruña, Spain
diego.alvareze@udc.es

Abstract

Deep learning has demostrated its usefulness in reaching top-level performance. However, inter-database generalization is still a broad of concern due to the aroused differences between local and external datasets' performances. In this work we explore different deep learning model's combination strategies applied to a multi-database case of study in the domain of sleep medicine. More specifically, three ensemble combination methods (namely max-voting, output averaging and weighted combination using the Nelder-Mead search) are analyzed in comparison to baseline methods (local models, database assembly approach) in a sleep staging inter-database generalization task.

1 Introduction

Neural networks development and the huge amount of available data have opened new possibilities in the field of features extraction and pattern recognition. However, it is still a broad of concern the models' capability of generalization, in which the "true" performance is given by one single local train-test dataset partition. This issue has been studied in the past by the authors in the context of inter-database generalization in the domain of sleep medicine [1]. More precisely, in the sleep staging task, which involves the manual classification of the patient's sleep activity, essential in the diagnosis of sleep disorders [2]. In order to reduce the complexity and time needed to carry out this task, several automatic sleep staging methods have been implemented [3]. Nevertheless, despite the promising performance reported, it was not satisfactory when dealing with external databases, although they addressed the same assignment [4]. This problem arises due to sources of variability which result in poor generation to external

*This study has been funded under project ED431H 2020/10 of Xunta de Galicia. Authors wish to acknowledge the support received from the Centro de Investigación de Galicia "CITIC", funded by Xunta de Galicia and the European Union (European Regional Development Fund-Galicia 2014-2020 Program), by grant ED431G 2019/01.

datasets. In order to deal with this situation, in a previous study we have proposed an approach based on ensemble combination of local deep learning models, showing an improvement in inter-database generalization performance [1]. This work embraces the same case of study with a focus on the investigation of different combination strategies. We explore three different ensemble methods, namely max-voting, output averaging, and weighted combination using the Nelder-Mead search optimization algorithm. We compare the performance that results from each of these strategies on a multiple-database scenario. In addition, the results are compared with the classical widespread approach to boost a learning model's generalization by increasing the amount of training data. Based on the results of our experimentation, we analyze and discuss the inconveniences and advantages of each of these approaches.

2 Materials and Methods

Six heterogeneous and independent open sleep databases were used: *Haaglanden Medisch Centrum sleep staging database (HMC), St. Vicent's Hospital/University College Dublin Sleep Apnea Database (Dublin), Sleep Health Heart Study (SHHS), Sleep Telemetry Study (Telemetry), DREAMS subject database (DREAMS)* and *ISRUC-SLEEP dataset (ISRUC)*. Each one contains sleep stages (wakefulness, N1, N2, N3 and R) corresponding to 30s time "epochs" segmentation of polysomnographic (PSG) recordings, comprising monitoring of electroencephalographic (EEG), electromyographic (EMG) and electrooculographic (EOG) activity [2]. All data were obtained from public online repositories and is digitally encoded using the open EDF(+) format [5]. A CNN-LSTM deep learning architecture is used based on an earlier work [1]. The model was re-implemented for this study using Python (version 3.9.7, Tensorflow 2.7.0). In order to take into account sleep staging sequence dependencies, a 5-epoch input length was used. More detailed description of the used databases, the PSG processing pipeline, and the CNN-LSTM architecture is available at [1].

Three experiments were designed. Each database is split into train (TR), validation (VAL) and test (TS) datasets, using a 80-20 proportion. *Experiment 1: Local models.* Six CNN-LSTM are built using the corresponding TR and VAL datasets. Each model is evaluated on its own TS dataset to measure "local" performance, and on the remaining databases using full content, intending to asses the corresponding external performance. *Experiment 2: Database-combined models.* We built six database-combined CNN-LSTM using a leave-one-out approach: five of the gathered databases are pooled to produce combined TR and VAL datasets by joining the TR and VAL datasets partitions of Experiment 1. Full data of the discarded database is used to evaluate the model. *Experiment 3: Ensemble models.* Six ensemble models are created using every possible combination of local models resulting from Experiment 1 using the same leave-one-out approach as in Experiment 2. Three ensemble strategies are tested. First, max-voting where the final classification corresponds to the most voted class. Second, output averaging where the output is calculated using the average of the five model's softmax normalized output class scores. And third, a weighted combination of the corresponding softmax normalized output class scores from each of the models integrating the ensemble. The Nelder-Mead search algorithm is used to find the optimum weight combination. A maximum of 10 iterations are allowed during the optimization. For all experiments we use Cohen's Kappa (κ) as the main metric of performance.

Table 1: Generalization performance (Cohen's Kappa)

CNN-LSTM

Test database	Local	Combined	Max-voting	Averaging	Nelder
ISRUC	0,51	0,66	0,55	0,57	0,57
SHHS	0,51	0,63	0,59	0,62	0,62
DREAM	0,51	0,71	0,57	0,62	0,63
Telemetry	0,46	0,67	0,58	0,59	0,60
HMC	0,43	0,59	0,50	0,53	0,54
Dublin	0,48	0,63	0,61	0,63	0,63
Average	0,49	0,65	0,57	0,59	0,60

3 Results

For each of the tested methods, the resulting performance per database is shown in Table I. Local models' generalization performance was obtained by averaging all the five out of six individual local models when the corresponding dataset is used in the external test scenario. According to our results, the worst prediction for local models was obtained for HMC (κ=0.43). Likewise, poor generalization using HMC can be seen for combined and ensemble models (κ=0.50-0.59), resulting in the most difficult database to be predicted. Nevertheless, differences between strategies are noticeable when considering the best prediction scenario. In this respect, the best scenario for local models is when evaluating ISRUC, SHHS and DREAMS, with associated averaged κ=0.51. DREAMS highlights for combined models (κ=0.71), whereas Dublin is best predicted by ensembles, regardless of the specific strategy (κ=0.61-0.63). Overall the best performance is achieved by combined models (κ=0.65) followed by Nelder-Mead (κ=0.60), and ensemble's output averaging (κ=0.59).

4 Discussion and conclusions

Data combination and ensemble methods can raise model's generalization robustness (Table I). Overall, combined models have achieved the best performance. This is expected since enlarging the amount and heterogeneity of training data is well-known to contribute to raise generalization in deep learning. However, this approach has several disadvantages. First, all data needs to be stored in a centralized dataset, which could be problematic, technically, and due to data confidentiality. Second, it is not possible to add new datasets without dealing with the so-called "catastrophic forgetting", leading to re-train the whole model again. By contrast, ensemble strategies combine independent local models without the need of re-train, therefore, it is possible to add new models dynamically through time. Furthermore, this approach overcomes the problem of sharing sensitive information, as it is the model and not the data what is shared for the construction of the ensemble. Note that, despite Nelder-Mead method has obtained the best performance, it uses the combined VAL dataset as reference for guiding the weight's optimization search. In contrast, this is not a problem for max-voting and output averaging strategies. Further investigation is needed in order to explore different deep learning strategies to improve models' scalability and inter-database generalization reliability.

References

[1] Diego Alvarez-Estevez and Roselyne M Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PloS one*, 16(8):e0256111, 2021.

[2] R.B. Berry, S.F. Quan, A Abreu, M Bibbs, L Del Rosso, et al. The aasm manual for the scoring of sleep and associated events: rules, terminology and technical specifications (version 2.6). *Darien, IL: American Academy of Sleep Medicine*, 2020.

[3] Luigi Fiorillo, Alessandro Puiatti, Michela Papandrea, Pietro-Luca Ratti, Paolo Favaro, Corinne Roth, Panagiotis Bargiotas, Claudio L. Bassetti, and Francesca D. Faraci. Automated sleep scoring: A review of the latest approaches. *Sleep Medicine Reviews*, 48:101204, December 2019.

[4] Diego Alvarez-Estevez and Isaac Fernández-Varela. Addressing database variability in learning from medical data: An ensemble-based approach using convolutional neural networks and a case of study applied to automatic sleep scoring. *Computers in Biology and Medicine*, 119:103697, 2020.

[5] B Kemp and J Olivan. European data format 'plus' (EDF+), an EDF alike standard format for the exchange of physiological data. *Clinical Neurophysiology*, 114:1755–1761, 2003.