Explainable Online Deep Neural Network Selection using Adaptive Saliency Maps for Time Series Forecasting - Supplementary Material

Amal Saadallah, Matthias Jakobs, and Katharina Morik

Artificial Intelligence Group
Department of Computer Science, TU Dortmund, Germany
{firstname.lastname}@tu-dortmund.de

1 Code and reproducibility

We provide all the Python code necessary for training the individual models and applying our OS-PGSM method and all its variants, including all datasets, under the following link: `https://github.com/MatthiasJakobs/os-pgsm`

In addition, the archive contains a README file, which aims to help in reproducing the results.

2 OS-PGSM algorithm

In addition to the pseudo code for our OS-PGSM algorithm, provided in the paper, we also show how it works in Figure 1. There, we show an example time series X_{ω}^{val} with a moving window approach of step size z. For each window of size n_{ω}, we compare the prediction of all models C_j from the model pool and choose the best performing one. Then, for this model, we compute the region of competences (shown in red) and add it to the corresponding RoC^j buffer.

3 Experiments

3.1 Datasets

All used datasets, together with a short description, can be found in 1. For the datasets from the UEA & UCR time series classification repositories, we "converted" them to time series forecasting datasets by taking the first feature vector X_0 from the respective training set and using that for splitting into training, validation and test parts. We also used a total of 80 datasets from the M4 competition [6]. More precisely, we used the first 20 columns of the hourly, monthly, weekly and daily tables given by the challenge. The extracted columns were cleaned by skipping the NaN values at the beginning and end of each time series, should they exist.

In total, we collected 102 datasets from diverse application fields, including audio, sensor values and financial data. All datasets were preprocessed by subtracting their mean and dividing by their standard deviation.

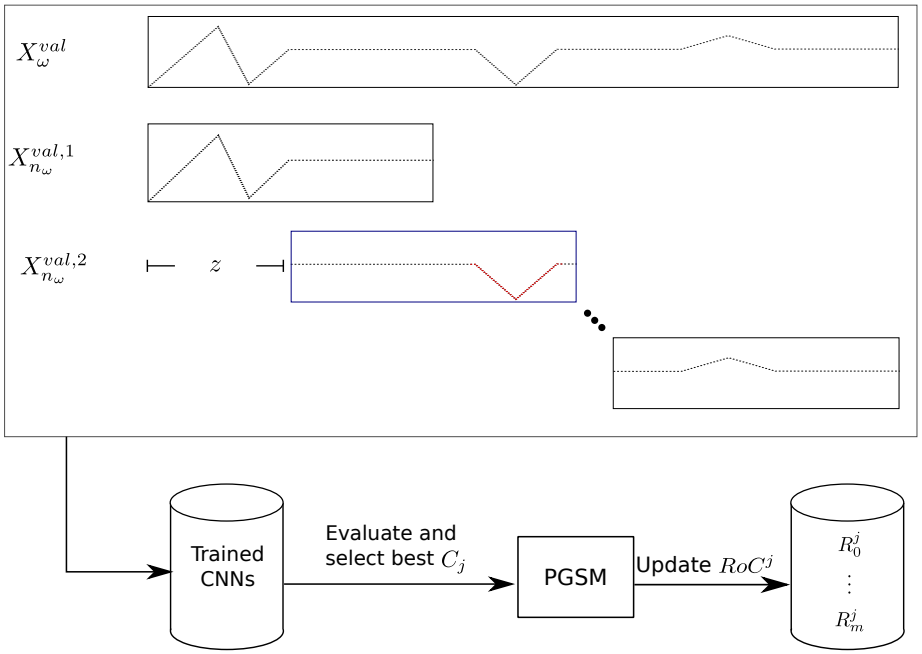

Fig. 1: Schematic visualization of our approach for extracting RoCs from the validation time series.

3.2 Rank plots

Figure 2 shows the ranks of our approaches in comparison to state-of-the-art methods and baselines as boxplots. Our methods outrank all comparison methods, including ADE-Single. Note that rank of 1 means the model was the best performing on all time series.

3.3 Critical Difference Diagram

To further investigate the differences in the average ranks, we use the post-hoc Bonferroni-Dunn test [3] to compute critical differences. We present the critical differences between the methods relative to each other in figure 3. We note critical differences between OS-PGSM and most of the other methods, with the exceptions of ADE-Single. However, unlike OS-PGSM ADE-Single does not share critical differences to the other methods.

3.4 Investigating the Regions of Competences RoCs

As can be seen in Figure 4, five single models do not have a region of competence after applying the OS-PGSM method on the AbnormalHeartbeat dataset. Notice that C^{11} and C^2 contain a lot more regions of confidence, which explain their

Name	Nr of time series	Source	Characteristics
Total rents	1		
Temperature	1	Bike sharing	Hourly, Jan. 1, to Mar. 01, 2011
Amount registered	1	[2]	
AbnormalHeartbeat	1		3053 measurements (4kHz)
CatsDogs	1		14773 audio samples (16kHz)
Cricket	1		1197 accelerometer readings (184Hz)
EOGHorizontalSignal	1		1250 measurements (1kHz)
EthanolConcentration	1	UEA	1 second spectrum measurement
Mallat	1	& UCR [1]	1024 measurements (simulated)
Phoneme	1		1024 samples of audio
PigAirwayPressure	1		2000 pressure measurements
Rock	1		2844 samples of spectrum analysis
SNP500	1		
NASDAQ	1		
DJI	1	UCI [4][5]	Daily closing, 2010 to 2017
NYSE	1		
RUSSELL	1		
Humidity RH1	1		
Humidity RH2	1	Appliances	10-minute steps, Jan. 11,
Temperature T4	1	Energy [4]	2016 17:00 to May 27, 2016 18:00
Temperature T5	1		
Avg. Cloud Cover	1	NREL [7]	Hourly, Apr. 25 to Aug. 25, 2016
M4 competition	80	[6]	Subset of 20 hourly, monthly, weekly and daily time-series each

Table 1: All used datasets in the experiment section of the paper. In total, 102 datasets were used.

performance as the two best candidate models in the evaluation. Also notice that the line strength of the RoCs corresponds with how many identical time series are in the models region of competence. For example, C^7 contains two distinct types of peaks, one of it is more darkly colored, indicating that there are actually numerous identical shapes/ patterns in the region of competence of one model.

References

1. Bagnall, A., Lines, J., Bostrom, A., Large, J., Keogh, E.: The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery **31**, 606–660 (2017)
2. Cerqueira, V., Torgo, L., Pinto, F., Soares, C.: Arbitrated ensemble for time series forecasting. In: Joint European conference on machine learning and knowledge discovery in databases. pp. 478–494. Springer (2017)
3. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research **7**, 1–30 (2006)
4. Dua, D., Graff, C.: UCI machine learning repository (2017), `http://archive.ics.uci.edu/ml`

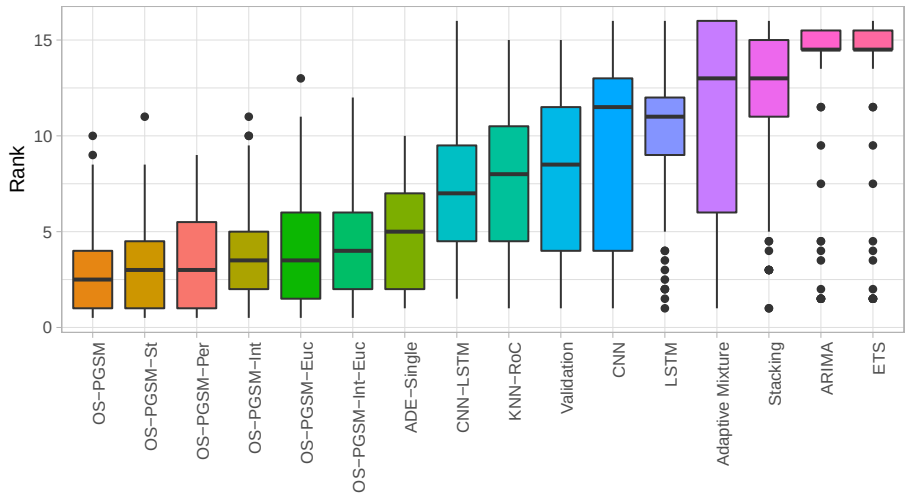

Fig. 2: Distribution of the ranks of the candidate models across the different time series.

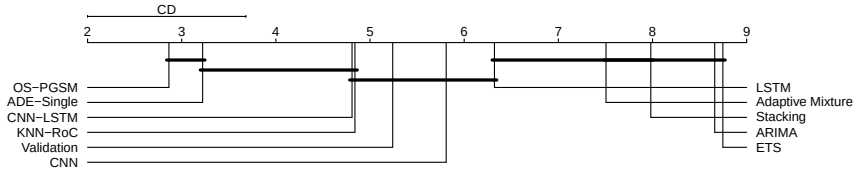

Fig. 3: Critical difference diagram for the post-hoc Bonferroni-Dunn test, comparing OS-PGSM with the other baseline ensemble methods.

5. Hoseinzade, E., Haratizadeh, S.: Cnnpred: Cnn-based stock market prediction using a diverse set of variables. Expert Systems with Applications **129**, 273–285 (2019). https://doi.org/https://doi.org/10.1016/j.eswa.2019.03.029, `https://www.sciencedirect.com/science/article/pii/S0957417419301915`
6. Makridakis, S., Spiliotis, E., Assimakopoulos, V.: The m4 competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting **36**(1), 54–74 (2020). https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.04.014, `https://www.sciencedirect.com/science/article/pii/S0169207019301128`, m4 Competition
7. Stoffel, T., Andreas, A.: Nrel solar radiation research laboratory (srrl): Baseline measurement system (bms); golden, colorado (data) (7 1981). https://doi.org/10.7799/1052221

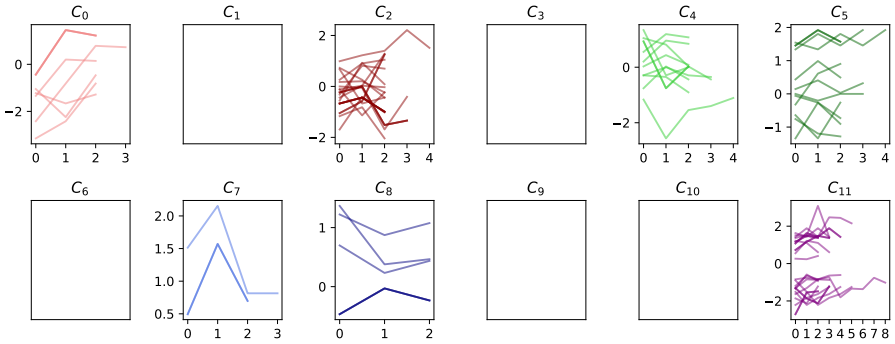

Fig. 4: Regions of competences for all forecasters after using the OS-PGSM-St method on the AbnormalHeartbeat dataset. Notice that some forecasters do not have a region of competence and are just shown empty.