LAMARR
INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Interpretable and Explainable Online Model Selection

Matthias Jakobs

June 17, 2024

technische universität dortmund
Fraunhofer IAIS
UNIVERSITÄT BONN
Fraunhofer IML

Problem Statement

- Online setting: Input data points arrive one by one
- Prediction using pretrained models (*offline*)
- **Problem:** Model performance can degrade with time
- $\Rightarrow$ Pool of pretrained models, of which we want to choose the best

Application scenario: Time Series Forecasting, but our frameworks are general

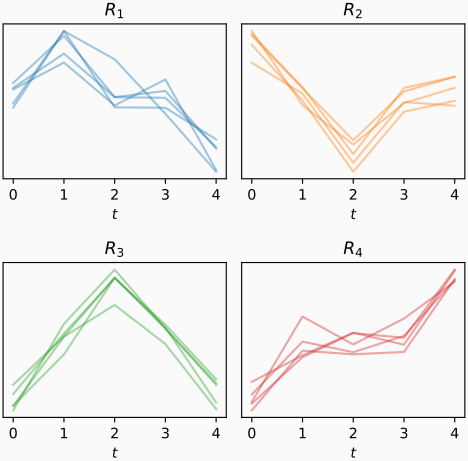
Region of Competence (RoC)

Framework for model competence

Each model in a pool remembers the types of input it excelled at.
At inference time: RoC serves as indicator for expected performance

Assumption:

Input close to RoC member
$\Rightarrow$ High performance



Concept drift

Real phenomenon in real applications

Adaption necessary to keep performance high

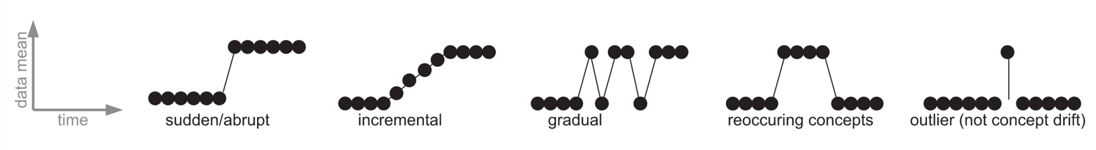


Figure 1: From [1]

TSMS [4]

Recently proposed methods utilize pools of **Neural Networks**

- Amal Saadallah, Matthias Jakobs, and Katharina Morik. **"Explainable online ensemble of deep neural network pruning for time series forecasting".** In: *Machine Learning* 111.9 (2022)
- Amal Saadallah, Matthias Jakobs, and Katharina Morik. **"Explainable Online Deep Neural Network Selection Using Adaptive Saliency Maps for Time Series Forecasting".** In: *Machine Learning and Knowledge Discovery in Databases. Research Track*. Ed. by Nuria Oliver et al. Cham: Springer International Publishing, 2021, pp. 404–420. ISBN: 978-3-030-86486-6

✓ High performance

✗ Low interpretability of base models

TSMS [4]

- Pool is made of tree-based models for improving interpretability
  - Decision Trees, Random Forests, Gradient Boosting Trees
- Refinement of RoC members using **Shapley values**
  - Computation efficient for tree-based models

TSMS [4]

Framework from cooperative Game Theory to distribute contributions to an outcome fairly unto the participants

Value function v, set of all players N

How much did player $i \in N$ contribute to value $v(N)$?

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \binom{|N|-1}{|S|}^{-1} \frac{v(S \cup \{i\}) - v(S)}{|N|}$$

TSMS [4]

Shapley values are the only attribution method satisfying the following axioms:

1. **Efficiency**: $\sum_{i \in N} \phi_i(v) = v(N) - v(\emptyset)$
2. **Null player**: $v(S \cup \{i\}) = v(S)\ \forall S \subseteq N \setminus \{i\} \Rightarrow \phi_i(v) = 0$
3. **Symmetry**: $v(S \cup \{i\}) = v(S \cup \{j\})\ \forall S \subset N \setminus \{i, j\} \Rightarrow \phi_i(v) = \phi_j(v)$
4. **Linearity**: $\phi_i(v + w) = \phi_i(v) + \phi_i(w)\ \forall i \in N$
    - Important for Random Forests

TSMS [4]

In Machine Learning, players are features and value functions are usually defined as

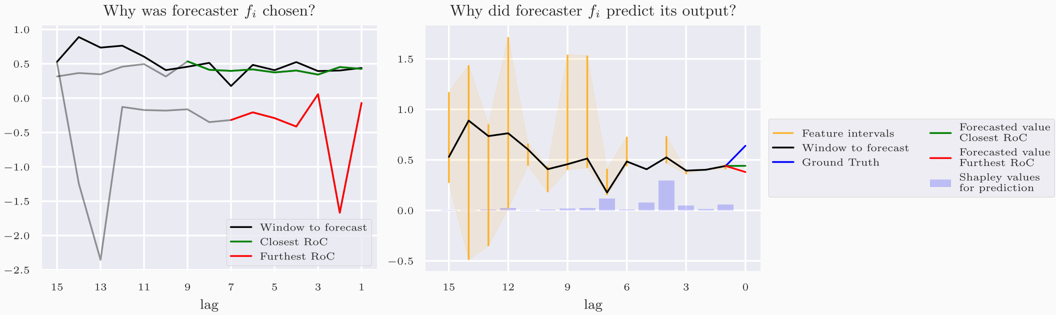
$$v_g(x, S) = \mathbb{E}_{X \sim \mathcal{X}}[g(X \mid X_S = x_S)]$$

for a given model prediction function g.

We want to explain the loss, changing v to

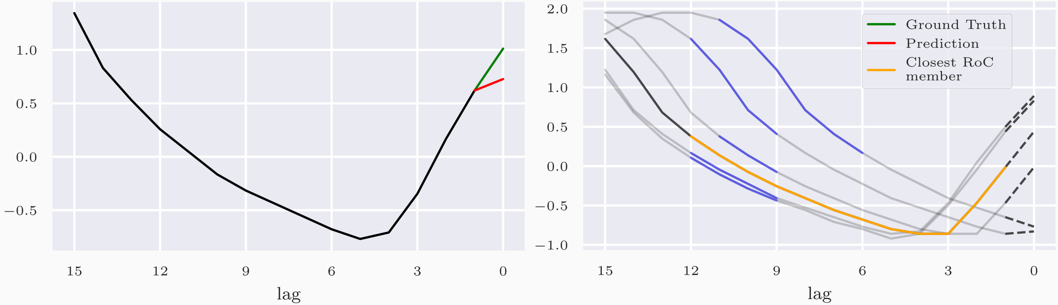
$$v_g(x, y, S) = \mathbb{E}_{X \sim \mathcal{X}}[(g(X \mid X_s = x_S) - y)^2]$$

TSMS [4]


Why was forecaster f_i chosen?
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
15
13
11
9
7
5
3
1
lag
Window to forecast
Closest RoC
Furthest RoC
Why did forecaster f_i predict its output?
1.5
1.0
0.5
0.0
−0.5
15
12
9
6
3
0
lag
Feature intervals
Window to forecast
Ground Truth
Forecasted value Closest RoC
Forecasted value Furthest RoC
Shapley values for prediction

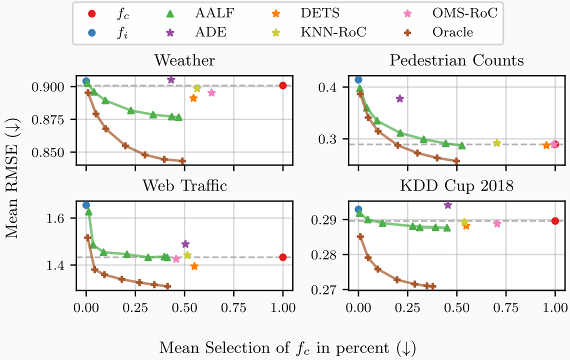
TSMS [4]

Visualization of RoC for model f_{17}



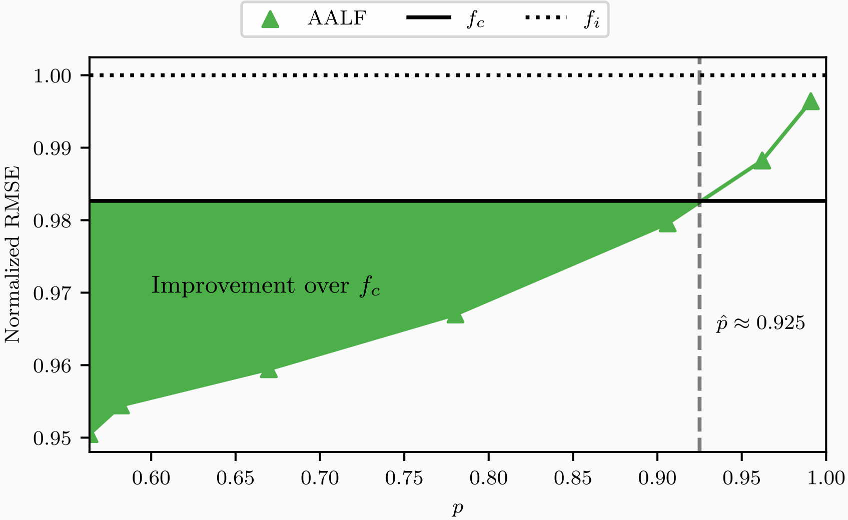
AALF *(under review)*

- Model pool restricted to two models
  - Deep Learning Ensemble (f_c)
  - Linear Regression (f_i)
- Use f_i whenever possible
  - Meta-learner decides based on historic data
- Hyperparameter p can be set to allow for more predictive performance or focus on more interpretability



$\Rightarrow$ Some predictions of f_i have very high error, but most are *good enough*

AALF *(under review)*



References

[1] João Gama et al. **"A survey on concept drift adaptation".** In: *ACM computing surveys (CSUR)* 46.4 (2014), pp. 1–37.

[2] Amal Saadallah, Matthias Jakobs, and Katharina Morik. **"Explainable online ensemble of deep neural network pruning for time series forecasting".** In: *Machine Learning* 111.9 (2022).

[3] Amal Saadallah, Matthias Jakobs, and Katharina Morik. **"Explainable Online Deep Neural Network Selection Using Adaptive Saliency Maps for Time Series Forecasting".** In: *Machine Learning and Knowledge Discovery in Databases. Research Track*. Ed. by Nuria Oliver et al. Cham: Springer International Publishing, 2021, pp. 404–420. ISBN: 978-3-030-86486-6.

[4] Matthias Jakobs and Amal Saadallah. **"Explainable Adaptive Tree-based Model Selection for Time-Series Forecasting".** In: *2023 IEEE International Conference on Data Mining (ICDM)*. 2023, pp. 180–189. DOI: *10.1109/ICDM58522.2023.00027*.