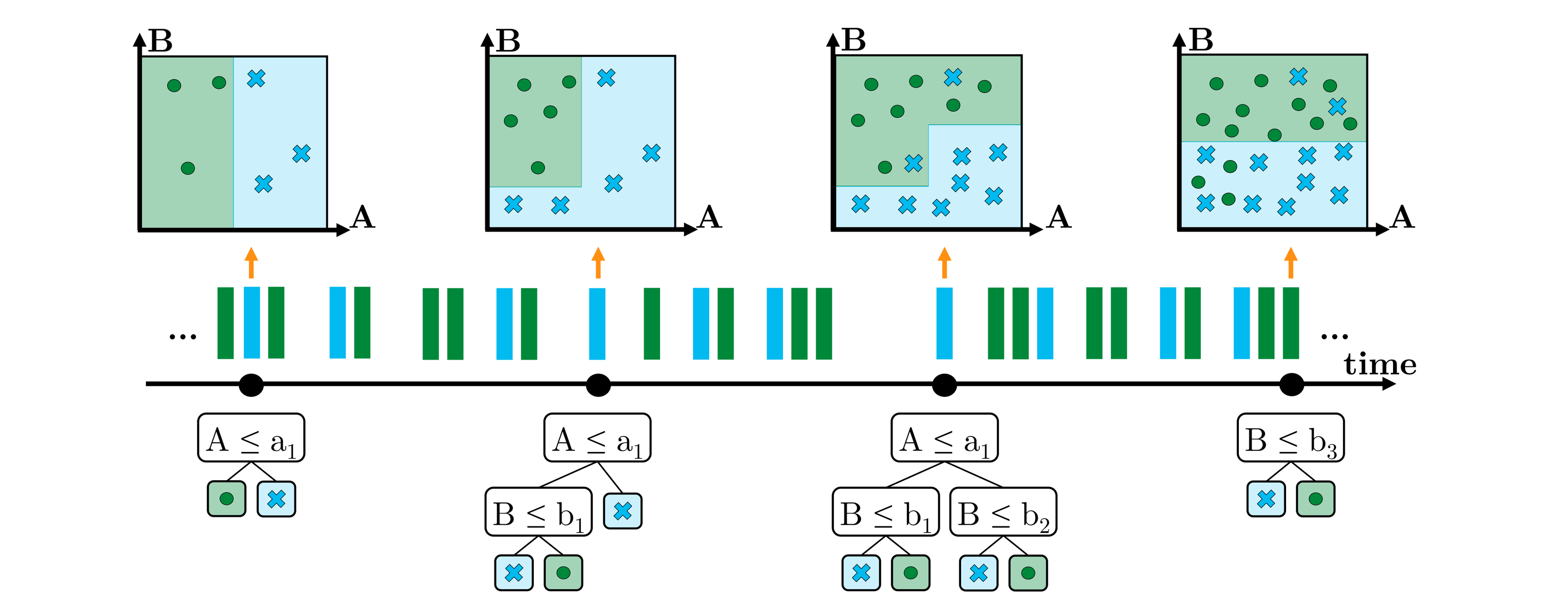
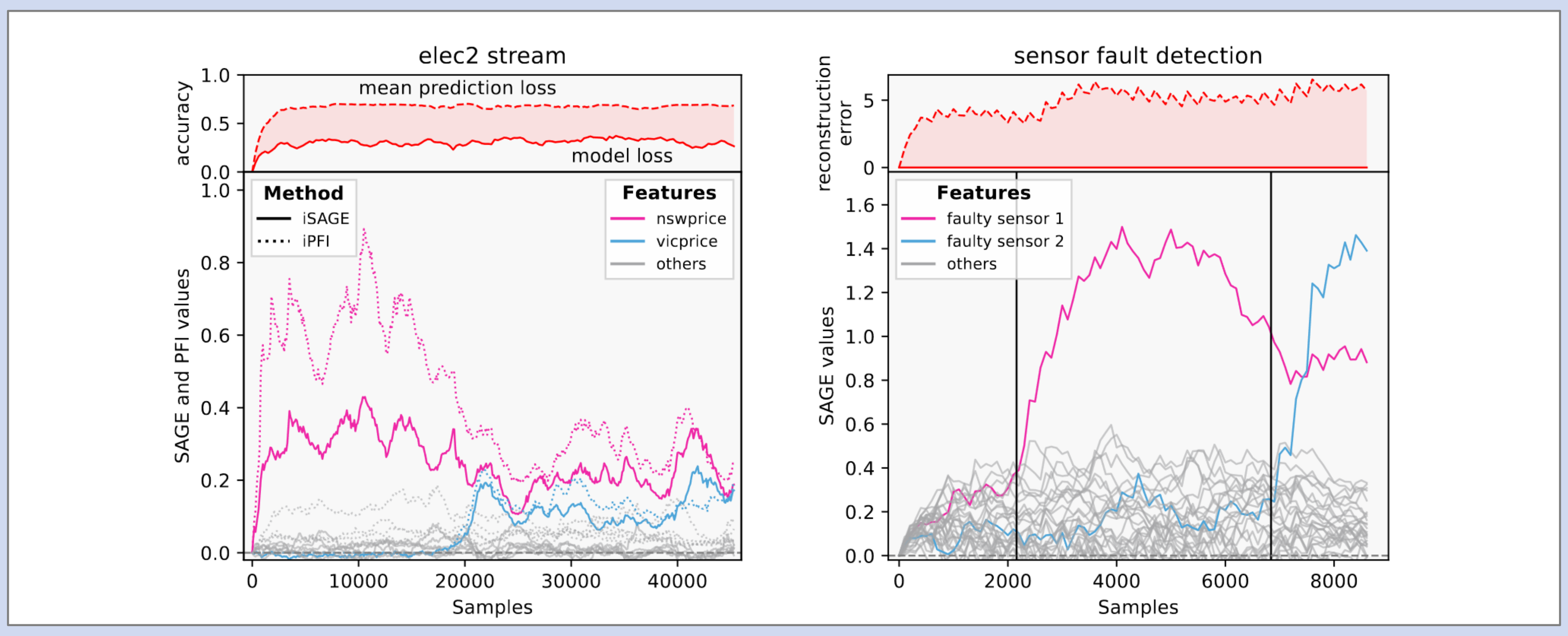
iSAGE: An Incremental Version of SAGE for Online Explanation on Data Streams

Maximilian Muschalik[1,*], Fabian Fumagalli[2,*], Barbara Hammer[2], Eyke Hüllermeier[1]

The Problem: Changing Black Box Models



A Solution: Incremental Model-Agnostic Global FI



SAGE (Shapley Additive Global Explanation)

SAGE Values are Shapley Values

(X,Y) data distribution on $\mathcal{X}\times\mathcal{Y}$ $f:\mathcal{X}\to\mathcal{Y}$ black box model $\ell:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$ loss function

Explanation Goal: Difference between Model Loss *with* Features and *without*

$$\nu(D) := \underbrace{\mathbb{E}_Y[\ell(\bar{y},Y)]}_{\text{no feature information}} - \underbrace{\mathbb{E}_{(X,Y)}[\ell(f(X),Y)]}_{\text{with feature information}} \quad \text{with mean prediction } \bar{y} := \mathbb{E}_X[f(X)]$$

Requirement: Restricted Improvement in Loss given $S\subset D$

$$\nu(S) := \mathbb{E}_Y[\ell(\bar{y},Y)] - \mathbb{E}_{(X,Y)}[\ell(f(X,S),Y)] \quad \text{with restricted model } f(x,S)$$

SAGE values ϕ of feature $i\in D$, i.e. Shapley values (Shapley 1953)

$$\phi(i) := \sum_{S\subset D\setminus\{i\}} \frac{1}{d}\binom{d-1}{|S|}^{-1}[\nu(S\cup\{i\})-\nu(S)]$$

Restricted Model

$\bar{S} := D\setminus S$

Observational SAGE

$$f^{\text{obs}}(x,S) := \mathbb{E}\left[f(x^{(S)},X^{(\bar{S})}) \mid X^{(S)}=x^{(S)}\right]$$

"true to the data"

Interventional SAGE

$$f^{\text{int}}(x,S) := \mathbb{E}\left[f(x^{(S)},X^{(\bar{S})})\right]$$

"true to the model"

In Practice: $\hat{f}(x,S) := \frac{1}{M}\sum_{m=1}^{M} f(x^{(S)},\tilde{x}_m^{(\bar{S})})$

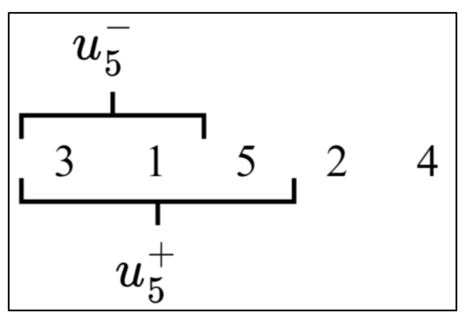
→ requires **sampling mechanisms** for replacements $\tilde{x}_m^{(\bar{S})}$

Computation with Permutation Sampling

$$\hat{\phi}^{\text{SAGE}}(i) := \frac{1}{N}\sum_{n=1}^{N} \underbrace{\ell(\hat{f}(x_n,u_i^-(\pi_n)),y_n) - \ell(\hat{f}(x_n,u_i^+(\pi_n)),y_n)}_{\Delta_n(i)}$$

Illustration of Permutation Sampling

Sample π_n :



$$\Delta_n(5) = \ell(\hat{f}(x_n,\{3,1\}),y_n) - \ell(\hat{f}(x_n,\{3,1,5\}),y_n)$$

Incremental SAGE (iSAGE)

iSAGE Estimator

The iSAGE estimator is **recursively** defined:

$$\text{iSAGE: } \hat{\phi}_t(i) = (1-\alpha)\cdot\hat{\phi}_{t-1}(i) + \alpha\cdot\Delta_t(i)$$

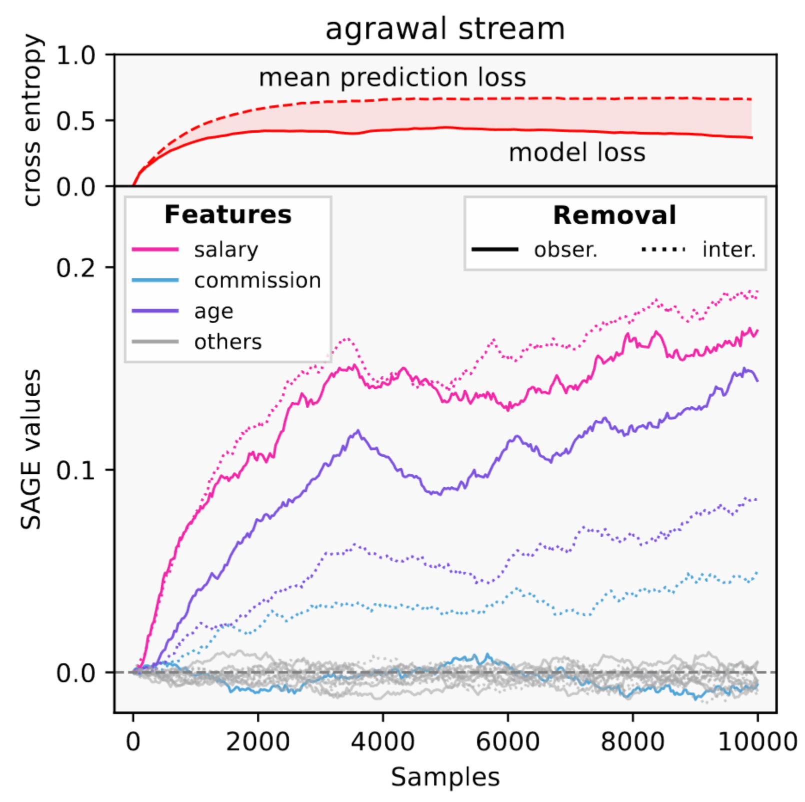
where $\alpha>0$ and computation starts at $0<t_0<t$ with $\hat{\phi}_{t_0-1}:=0$

Updates are computed **incrementally** for each time point

$$\Delta_t(i) := \ell(\hat{f}_t(x_t,u_i^-(\pi_t)),y_t) - \ell(\hat{f}_t(x_t,u_i^+(\pi_t)),y_t)$$

→ combining $\Delta_t(i)$ over time requires efficient online **sampling** mechanisms for the **replacement** values

Observational and Interventional iSAGE



Setting:
- $X^{\text{com.}}$ depends on X^{salary}
- knowledge about X^{salary} allows perfect reconstruction of $X^{\text{com.}}$
- $X^{\text{com.}}$ should not be important

observational and **interventional** iSAGE retrieve **different** FI scores
- observational iSAGE shows that $X^{\text{com.}}$ is not important
- interventional iSAGE shows that the model has learned to use $X^{\text{com.}}$ (i.e. decision splits exist for $X^{\text{com.}}$)

Theoretical Guarantees

Assumptions: static model $f_t\equiv f$ and data generating process $(X_t,Y_t)\equiv(X,Y)$

Theorem (Convergence)

For iSAGE $\hat{\phi}_t(i)\to\phi_t(i)$ for $M\to\infty$ and $t\to\infty$.

Theorem (Variance)

The variance of iSAGE is controlled by α, i.e. $\mathbb{V}[\hat{\phi}_t(i)]=\mathcal{O}(\alpha)$.

Theorem (Confidence Bounds)

Given the SAGE estimator $\hat{\phi}_t^{SAGE}(i)$ computed at time t over all previously observed data points, it holds for iSAGE with $M\to\infty$, $\alpha=\frac{1}{t}$ and every $\epsilon>(1-\alpha)^{t-t_0+1}$ that

$$\mathbb{P}\left(|\hat{\phi}_t(i)-\hat{\phi}_t^{SAGE}(i)|>\epsilon\right)=\mathcal{O}(\tfrac{1}{t}).$$

Open Source Implementation: iXAI

- works natively with **riverml.xyz**
- incorporates: iSAGE, iPFI, iPDP, and MDI
- looking for **collaborators**!

References

Castro, Javier, Daniel Gómez, and Juan Tejada (2009). "Polynomial calculation of the Shapley value based on sampling". In: Computers & Operations Research 36.5, pp. 1726–1730.
Covert, I., Lundberg, S. M., & Lee, S.-I. (2020). Understanding Global Feature Contributions With Additive Importance Measures. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020 (NeurIPS 2020).
Covert, Ian, Scott M. Lundberg, and Su-In Lee (2021). "Explaining by Removing: A Unified Framework for Model Explanation". In: Journal of Machine Learning Research 22.209, pp. 1–90.
Gama, João et al. (2014). "A Survey on Concept Drift Adaptation". In: ACM Computing Surveys 46.4, 44:1–44:37.
Shapley, L. S. (1953). "A Value for n-Person Games". In: Contributions to the Theory of Games (AM-28), Volume II. Princeton University Press, pp. 307–318.
Vitter, J. S. (1985). Random Sampling with a Reservoir. ACM Transactions on Mathematical Software, 11(1), 37-57.

Images in "SAGE Restricted Model" taken with permission from Kristin Blesh's XAI'23 presentation on "Unfooling SHAP and SAGE: Knockoff Imputation for Shapley Values"

* denotes equal contribution

1) LMU Munich, Munich, D-80539, Germany
2) Bielefeld University, Bielefeld, D-33594, Germany

Funded by Deutsche Forschungsgemeinschaft German Research Foundation